Forensic Statistics of Lineage DNA Markers

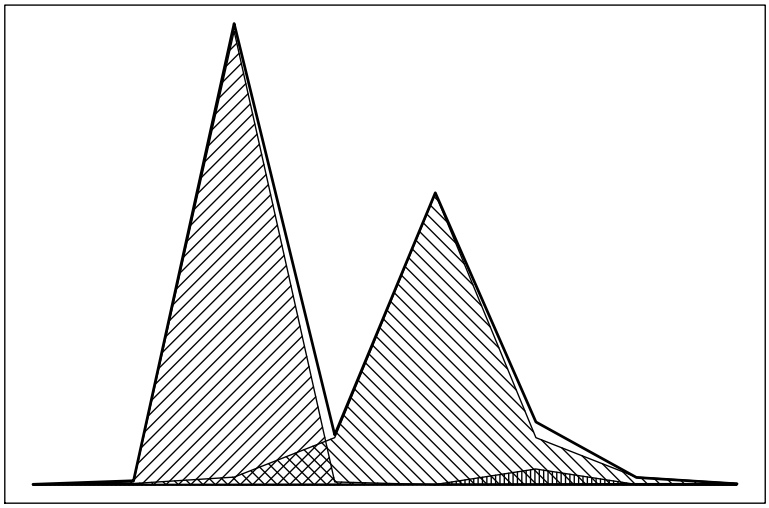
Andersen, Mikkel Meyer

Publication date:
2014

Document Version
Accepted author manuscript, peer reviewed version

PhD Report No. 25, 2014

Forensic Statistics of Lineage DNA Markers



Mikkel Meyer Andersen

AALBORG UNIVERSITY
Department of Mathematical Sciences

Forensic Statistics of Lineage DNA Markers

Mikkel Meyer Andersen

Thesis submitted: 9 Dec 2013

Thesis defended: 28 Feb 2014

PhD degree conferred: 30 Apr 2014

PhD supervisor: Poul Svante Eriksen
Department of Mathematical Sciences
Aalborg University, Denmark

PhD committee: Prof. Thore Egeland
Department of Chemistry, Biotechnology and Food Science
Norwegian University of Life Sciences, Norway

Prof. Lutz Roewer
Institute of Legal Medicine and Forensic Sciences
Charité Universitätsmedizin Berlin, Germany

Prof. Rasmus Waagepetersen
Department of Mathematical Sciences
Aalborg University, Denmark

DEPARTMENT OF MATHEMATICAL SCIENCES
Fredrik Bajers Vej 7G
9220 Aalborg, Denmark
http://www.math.aau.dk

ISSN 1601-8346

Forensic Statistics of Lineage DNA Markers

PhD Thesis
December 2013

Mikkel Meyer Andersen

Department of Mathematical Sciences
Aalborg University
Denmark

Preface

Revised version

Some minor corrections have been made to this revised version of the PhD thesis:

- Paper V: Minor typographical corrections.
- Paper VI: Corrections regarding moments of certain random variables (e.g. $\mathbf{E}[D]$ changed to $\mathbf{E}[|D|]$) that are also published as a corrigendum to the published paper.
- Paper IX: Added appendix with `R` implementation.

Mikkel Meyer Andersen
Aalborg, Denmark

Preface

This thesis summarises research work carried out during my employment as a PhD student at Department of Mathematical Sciences, Aalborg University, Denmark. The research was carried out in very close collaboration with Section of Forensic Genetics, Department of Forensic Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Denmark. A part of the work was carried out while based at Institute of Medical Informatics and Statistics, University of Kiel, Kiel, Germany.

The thesis is about statistical modelling of lineage DNA markers. The thesis consists of nine papers. Three of the papers have been published in widely recognised peer-reviewed journals, one in a journal's supplement series, one is a letter to the editor of a journal, one is submitted to a peer-reviewed journal, two are publically available preprints and one is in preparation for publication.

Each paper is self-contained with separate numbering of sections, figures, equations and bibliographies. At the very end of the thesis, the complete bibliography is provided.

In addition to the papers included in the thesis, four freely available open source packages for the statistical software, `R`, have been developed as part of the PhD work. Those packages are mentioned in the thesis as they implement some of the statistical methods described.

The first chapter is an introduction to the basic terminology and a brief recap of the role of lineage DNA markers in population and forensic genetics. Then, an outline of the remainder of the thesis is given. The outline gives a description of each paper that is less technical than the more concise abstracts that accompany the papers. In the last chapter of the thesis, topics for future research are briefly described.

Acknowledgements

Being my supervisor since 2009 during my MSc studies, Poul Svante Eriksen has been of great importance and inspiration to me. I want to express my deepest gratitude to him. Reasons include his suggested solutions to many of the problems I have worked on, his patience (especially when reading all my manuscript drafts), for encouraging me, his constant desire to help out and his willigness to share his extensive knowledge. Thank you heaps, Svante.

A lot of gratitude also goes to Niels Morling, who introduced me to the field of forensic genetics and has taken good care of me ever since. He has a clever eye that tends to quickly spot interesting projects and that can read a manuscript draft like no one else.

Sometimes Torben Tvedebrink must have thought of me as an annoying little brother always close to his heels. Especially at the first couple of conferences, where Torben used a lot of time introducing me to people within forensic genetics and forensic statistics. But not even once did he let me get such an impression. Thanks for taking such good care of me, Torben, and for many good and funny nights out in various countries.

Helle Smidt Mogensen has a deep insight into biochemistry and how forensic

genetics is actually exercised in the everyday lab life. This has been crucial when trying to create statistical models. Thanks for always explaining the complicated stuff over and over again with patience, Helle.

I also want to thank Jill Katharina Olofsson, Jeppe Dyrberg Andersen, Peter Johansen and Martin Mikkelsen for both professional discussions and lots of fun.

Søren Højsgaard has enlarged my interest for and helped develop my competences in statistical computing. Thank you for that.

Furthermore, I would like to thank both the entire Department of Mathematical Sciences at Aalborg University and the entire Section of Forensic Genetics at University of Copenhagen for always being friendly and helpful.

When I visited Michael Krawczak and all his coworkers in Kiel, Germany, they all made sure that I was taken care of and had three great months. Thank you all for that. I also want to thank all the people at the Hindenburgufer Guesthouse for great nights with all sorts of Georgian, Argentine, American, Portuguese, Russian, Greek and Egyptian food, drinks, music and culture bouts. Those months were exceptional.

I also thank Anders Gorst-Rasmussen for providing me with the LaTeX layout for this thesis.

Finally, thanks to my parents for fostering my love for learning. Thanks to my family and friends for their support. Most of all, I want to thank my wife, Rikke, for your love and tolerance. You have supported and encouraged me in all sorts of ways through these years. I am still grateful that you joined me for six truly amazing months in Australia during my MSc studies. That was a magnificent trip. Thank you.

Mikkel Meyer Andersen
Aalborg, Denmark

Summary

Forensic genetics utilising DNA information has shown to be invaluable in forensic investigations such as criminal, paternity and immigration cases.

DNA information is great to exclude a suspect in a crime case: If the DNA profile found at the crime scene does not match that of the suspect, the suspect is immediately exonerated. On the other hand, if the suspect's DNA profile matches that found at the crime scene, this evidence must be weighted to interpret the match correctly. This evidential weight is essential as a forensic genetic DNA profile is usually only a subset of the entire genome, hence people can have identical DNA profiles without having identical genomes. If the DNA profile found at the crime scene is very common in the population of interest, the evidential weight is not as large as if the DNA profile is very rare.

Lineage DNA profiles are DNA profiles that consist of markers on the Y chromosome (inherited as a unit through the paternal lineage) and on the mitochondrial DNA (inherited as a unit through the maternal lineage). In a number of crime cases, lineage DNA profiles are particularly helpful. DNA markers on the Y chromosome can help resolve cases when there is male/female cell admixture as for example in sexual assault cases. In such cases, it is possible to type only the Y chromosomal DNA profile and compare it to that of a male suspect.

Traditional DNA profiles are obtained from markers on the chromosomes in the cell nucleus. In some cases, the cell nucleus is destroyed or so deteriorated that a DNA profile cannot be made. This is for example often the case for hair shafts and very old biological samples. In such cases, it is often possible to obtain a DNA profile from the mitochondrial DNA as mitochondrial are more hardy and numerous than the cell nucleus.

Because lineage DNA markers have unique inheritance properties, these are also very interesting in population genetics because Y chromosome and mitochondrial DNA reflect male and female inheritance, respectively.

If a suspect's DNA profile matches that found at the crime scene, the weight of the evidence must be evaluated. To evaluate a match, statistical models for forensic genetics must be used. A lot of work on statistical models for interpreting traditional (non-lineage) DNA profiles have already been done and evaluation of the weight of such evidence is now routine work.

That is not the case for lineage DNA markers as the inheritance pattern means that the statistical methods for traditional DNA markers do not hold. In this thesis, the focus is on developing statistical models for lineage DNA markers as they are very different from traditional (non-lineage) DNA markers due to the inheritance patterns.

The main focus of this thesis is on estimating population frequencies of DNA profiles based on lineage DNA markers because this is essential in evaluating the evidential weight of a match. The main theories I have used for this part are that of Fisher-Wright populations, coalescent theory and finite mixtures of exponential families (a certain class of probability distributions). Cluster analysis methods have also been developed based on properties of the finite

mixture models.

A minor part of this thesis is on how to model errors that may arise during the process of obtaining a DNA profile from a biological trace. This process involves both chemicals and apparatus that can introduce errors and it is very important to understand the nature of such errors.

The main results of the thesis is that modelling of Y chromosomal short tandem repeat (Y-STR) DNA profiles is done well by a finite mixture of discrete Laplace distributions ('the discrete Laplace method'). Both inference of DNA profile frequencies and cluster analysis using this method (which has been implemented in publicly available open source software) yield state of the art results.

The thesis mainly deals with modelling the distribution of Y chromosomal short tandem repeat (Y-STR) DNA profiles, but as many of the statistical considerations are similar for other types of lineage DNA markers, I will refer to lineage DNA markers as a whole, especially in the introduction and epilogue. Concluding the thesis, I discuss how the obtained knowledge can be used in modelling other lineage DNA markers such as mitochondrial DNA and Y chromosomal single nucleotide polymorphism (Y-SNP).

Dansk resumé (Summary in Danish)

Retsgenetik, der udnytter DNA-information, har vist sig at være uvurderlig i retsmedicinske undersøgelser som eksempelvis straffe-, faderskabs- og immigrationssager.

DNA-information er fantastisk til at ekskludere en mistænkt i en straffesag: Hvis DNA-profilen fundet på gerningsstedet ikke matcher den mistænktes DNA-profil, kan DNA'et ikke stamme fra den mistænkte. Omvendt, hvis en mistænkts DNA-profil matcher den, der er fundet på gerningsstedet, skal dette bevismateriale vægtes for at kunne fortolke matchet korrekt. Den bevismæssige vægt er essentiel, da en retsgenetisk DNA-profil normalt kun er en delmængde af hele genomet, og dermed kan personer have identiske DNA-profiler uden at have identiske genomer. Hvis gerningsstedets DNA-profil er ofte forekommende i den relevante befolkning, er den bevismæssige vægt ikke så stor, som hvis DNA-profilen er meget sjælden.

DNA-slægtsprofiler er DNA-profiler, der består af slægtsmarkører, hvilket er DNA-markører på Y-kromosomet (der nedarves som en samlet enhed gennem faderslægten) og på det mitokondrielle DNA (der nedarves som en samlet enhed gennem moderslægten). I nogle straffesager er slægtsmarkører specielt brugbare. Slægtsmarkører på Y-kromosomet kan anvendes i sager, hvor der er blanding af celler fra mænd og kvinder som for eksempel i voldtægtssager. I sådanne sager er det muligt at lave en DNA-profil baseret udelukkende på DNA-markører på Y-kromosomet og sammenholde den med en tilsvarende fra den mistænkte.

Traditionelle DNA-profiler er baseret på DNA-markører på autosomale kromosomer i cellekernen. I nogle sager er cellekernerne ødelagt eller så nedbrudte, at det ikke er muligt at lave en traditionel DNA-profil. Dette er ofte tilfældet ved hårskafter (den del af håret, der er tilbage, når man fjerner hårroden) eller meget gamle biologiske prøver. I sådanne sager er det ofte muligt at lave en DNA-profil baseret på det mitokondrielle DNA, da mitokondrier er mere hårdføre og talrige end cellekerner.

Idet slægtsmarkører har unikke nedarvningsegenskaber, er de også meget interessante i populationsgenetik, da de afspejler faderslægten (Y-kromosom) og moderslægten (mitokondrie).

Hvis en mistænkts DNA-profil matcher DNA-profilen fra gerningsstedet, skal den bevismæssige vægt findes. For at gøre dette, anvender man statistiske modeller. Der er allerede forsket meget i statistiske modeller til tolkning af traditionelle DNA-profiler, og i dag er disse metoder blevet modnet tilstrækkeligt til at de kan anvendes rutinemæssigt i sagsarbejde.

Det er dog ikke tilfældet for DNA-slægtsprofiler: Nedarvningsegenskaberne betyder, at antagelserne i de statistiske modeller til tolkning af traditionelle DNA-profiler ikke længere er opfyldt. Derfor skal der anvendes anderledes statistiske modeller. Afhandlingens fokus er udvikling af sådanne statistiske modeller til tolkning af DNA-slægtsprofiler.

Et centralt punkt i dette er at estimere DNA-slægtsprofilers populationsfrekvenser. I afhandlingen er følgende teorier bl.a. anvendt: Fisher-Wright-populationer, coalescent-teori og endelige miksturer af eksponentielle familier

(en bestemt klasse af sandsynlighedsfordelinger). Metoder til at udføre klusteranalyse baseret på egenskaber for endelige miksturer er også blevet udviklet.

En mindre del af afhandlingen er modellering af fejl, der kan opstå under udvindingen af en DNA-profil fra biologisk materiale. Denne udvinding består både af kemikalier og apparatur, der kan forårsage fejl. Det er essentielt at forstå sådanne fejl for at kunne anvende DNA-profiler korrekt.

Hovedresultatet i denne afhandling er, at modelling af Y-kromosomale DNA-profiler baseret på *short tandem repeat* (STR) markører kan ske ved hjælp af en endelig mikstur af diskrete Laplace sandsynlighedsfordelinger. Både inferens af populationsfrekvenser og klusteranalyse ved hjælp af denne metode giver *state of the art* resultater. Metoden er blevet implementeret i offentligt tilgængeligt *open source*-software.

Afhandlingen drejer sig hovedsageligt om Y-kromosomale STR DNA-profiler, men mange af de statistiske overvejelser er tilsvarende for andre typer af slægtsmarkører. Derfor vil jeg omtale DNA-slægtsprofiler mere generelt, specielt i introduktionen og afslutningen.

I slutningen af afhandlingen er der en kort diskussion af, hvordan den opnåede viden kan bruges til at modellere andre typer slægtsmarkører baseret eksempelvis på mitokondrielt DNA og Y-kromosomal *single nucleotide polymophism* (Y-SNP).

List of papers

Thesis title: Forensic Statistics of Lineage DNA Markers
Name of PhD student: Mikkel Meyer Andersen
Supervisor: Poul Svante Eriksen, associate professor

Paper I. Andersen MM, Olofsson JK, Mogensen HS, Eriksen PS, Morling N (2011). *Estimating stutter rates for Y-STR alleles.* Forensic Science International: Genetics Supplement Series; **3**(1):e192-e193.

Paper II. Olofsson JK, Andersen MM, Mogensen HS, Eriksen PS, Morling N (2012). *Sequence variants of allele 22 and 23 of DYS635 causing different stutter rates.* Forensic Science International: Genetics. Letter to editor; **6**(6):e161-e162.

Paper III. Andersen MM, Mogensen HS, Eriksen PS, Olofsson JK, Asplund M, Morling N (2013). *Estimating Y-STR allelic drop-out rates and adjusting for interlocus balances.* Forensic Science International: Genetics; **7**(3):327-336.

Paper IV. Andersen MM, Caliebe A, Jochens A, Willuweit S, Krawczak M (2013). *Estimating trace-suspect match probabilities for singleton Y-STR haplotypes using coalescent theory.* Forensic Science International: Genetics; **7**(2):264-271.

Paper V. Andersen MM and Eriksen PS (2012). *Efficient forward simulation of Fisher-Wright populations with stochastic population size and neutral single step mutations.* arXiv: 1210.1773.

Paper VI. Andersen MM, Eriksen PS, Morling N (2013). *The discrete Laplace exponential family and estimation of Y-STR haplotype frequencies.* Journal of Theoretical Biology; **329**:39-51.

Paper VII. Andersen MM, Eriksen PS, Morling N (2013). *A gentle introduction to the discrete Laplace method for estimating Y-STR haplotype frequencies.* arXiv: 1304.2129.

Paper VIII. Andersen MM, Eriksen PS, Morling N. *Cluster analysis of European Y-chromosomal STR haplotypes using discrete Laplace distributions.* Submitted to Forensic Science International: Genetics (2013).

Paper IX. Andersen MM, Eriksen PS (2013). *Efficient iteratively reweighted least squares for weighted two-way analysis of variance.* In preparation.

This thesis has been submitted for assessment in partial fulfillment of the PhD degree. The thesis is based on the submitted or published scientific papers, which are listed above. Parts of the papers are used directly or indirectly in the extended summary of the thesis. As part of the assessment, co-author statements have been made available to the assessment committee and are also available at the Faculty. The thesis is not in its present form acceptable for open publication but only in limited and closed circulation as copyright may not be ensured.

Contents

Introduction

This introduction consists of three parts: (1) A recap of the basic terminology, (2) description of the role of lineage DNA markers in genetics and (3) descriptions of each chapter that are less technical than the more concise abstracts that accompany the papers.

1. Terminology

First, a recap of the basic terminology is given. Please, refer to the books by Butler (2001, 2005, 2010, 2012) for details.

The DNA markers most often discussed in the thesis are short tandem repeat (STR) markers. An STR is a repeated sequence of 2-6 nucleotides where the number of repeats is called the allele. The allele is the quantity of interest for identification as it varies between individuals. A DNA marker is then the allele (the number of repeats) at a particular position in the genome. The position in the genome is called the locus (the plural of locus is loci).

Lineage DNA markers are DNA markers on either the Y chromosome or the mitochondrial DNA. Because both the Y chromosome and the mitochondrion is inherited as a unit from the father and mother, respectively, lineage DNA markers constitute a DNA profile that is called a haplotype (after the Greek word for onefold and was first used by Piazza *et al.* (1969)). This is in contrast to traditional DNA markers on the autosomes (the 22 pairs of non-sex chromosomes). Here, two values (e.g. two alleles) for each locus are obtained (one from the mother and one from the father), and the source of the values cannot be infered from just the DNA profile. Also, the loci in a traditional DNA profile are assumed statistically independently because they are taken from various chromosome pairs and due to recombination between loci on the same chromosome pair. This is not the case for lineage DNA profiles because all loci are inherited as a haplotype, i.e. as a unit. This means that the statistical properties for lineage DNA markers are widely different from those of traditional DNA markers on the autosomes.

2. Lineage DNA markers in genetics

Lineage DNA markers on the Y chromosome or mitochondrial DNA (mtDNA) are of great interest to both forensic and population genetics due to the patrilineal inheritance of the Y chromosome and matrilineal inheritance of the mtDNA.

In forensic genetics, Y chromosomal markers can be used when the interest is in analysing male DNA that is masked by large amounts of female DNA as described by Gill *et al.* (1985); Sibille *et al.* (2002); Roewer (2009). In some forensic settings, the biological material is in poor condition such that no or only a few cell nuclei are present making the DNA from the chromosomes impossible to extract. This is e.g. the case with very old samples and hair shaft samples. In such cases, mtDNA can sometimes be extracted as described by Sullivan *et al.* (1991) and sequenced. Hence, lineage DNA markers are important in forensic

genetics as they help to solve cases that are otherwise difficult or even impossible to investigate using traditional methods.

Because lineage DNA markers have unique inheritance properties, these are also very interesting in population genetics because Y chromosome and mtDNA reflect male and female inheritance, respectively. Cann *et al.* (1987) demonstrated how mtDNA could be used for population genetic studies and Roewer *et al.* (2005) demonstrated how Y-STR markers could be used to infer recent historical events in the European Y-STR haplotype distribution.

3. Outline

3.1. *Error modelling*

The process of obtaining a DNA profile from biological material involves both chemicals and apparatus that can introduce errors. To interpret the resulting DNA profile correctly, it is essential to understand the error phenomena.

One of the very important biochemical techniques in constructing a DNA profile is the polymerase chain reaction (PCR). The PCR method amplifies a few copies of DNA to thousands or even millions of copies. It is described in more detail by Butler (2001, 2005, 2010, 2012). The Nobel Prize in Chemistry in 1993 was awarded to Kary B. Mullis and Michael Smith for inventing the PCR method. Please, refer to `http://www.nobelprize.org/nobel_prizes/chemistry/laureates/1993/` for more details.

Paper I & II. *'Estimating stutter rates of Y-STR alleles'*

This chapter is based on Andersen *et al.* (2011); Olofsson *et al.* (2012).

During the PCR process, amplification DNA products that are one repeat unit shorter than the original allele arise. Less commonly, it also happens that longer products and even products with more than one repeat unit in difference are produced. These incorrect products are called stutters and are described in more detail by Butler (2001, 2005, 2010, 2012). Stutters will be amplified later in the process. This means that the end result will typically consist of a majority of the allele of the original DNA material and stutter artefacts. Because the errors are stochastic, the fraction of stutters in the end result is stochastic. To get an impression of the fraction that is normally observed – so that the result can be correctly interpreted – a statistical model must be used.

In this compilation of paper I, *'Estimating stutter rates for Y-STR alleles'*, and paper II, *'Sequence variants of allele 22 and 23 of DYS635 causing different stutter rates'*, a linear regression model was used. In this way, it was possible to obtain knowledge about the fraction of stutters. This can for example be used to detect mixtures of biological material (from e.g. two males) as only the sum of their DNA profiles can be observed and what looks like two alleles of an unbalanced mixture may actually be a stutter and an allele. This is important as mixtures call for distinct interpretation.

Paper III. *'Estimating Y-STR allelic drop-out rates and adjusting for interlocus balances'*

This chapter is based on Andersen *et al.* (2013d).

Some alleles are not detected because they have a mutation in what is called the primer binding site (a DNA anchor located next to the repeated sequence of 2-6 nucleotides). These alleles are called null alleles or silent alleles.

Another reason for alleles not showing up in the end product is so-called allelic drop-out. During the PCR process, amplification of the allele at a locus can fail such that the end product contains no allele. This can happen if the amount of input DNA is low or if the DNA is damaged, e.g. due to degradation. It can also happen with a large amount of healthy DNA, but the probability is much lower because the PCR process must fail simultaneous at several DNA fragments. This is opposed to null alleles, as they occur independently of the amount of DNA.

Again, the errors causing drop-outs are stochastic and a statistical model must be used to estimate the probability of a drop-out. To model this drop-out phenomenon, a logistic regression model was used together with inference in truncated normal distributions.

The drop-out probability is essential, especially in samples with low amounts of DNA.

3.2. *Haplotype distribution modelling*

In forensic genetics, it is often necessary to compare the plausibility of two case-relevant hypotheses on the basis of genetic data. The most consistent (and therefore generally recommended) way of doing so is to quantify the evidential weight by means of the likelihood ratio (e.g. Evett and Weir (1998)). Calculating the likelihood ratio in forensic case work is usually tantamount to quantifying the match probability between two genetic profiles under two different assumptions. One particularly important match probability in this context is the probability that a certain individual (e.g. the donor of a trace found at a crime scene) has the same DNA profile as another individual (usually a suspect) chosen randomly from the same population. If the trace haplotype is very common in the population, the evidential weight is not as large as if the trace haplotype is very rare in the population.

Let E be the evidence. The likelihood ratio quantifying the weight of the evidence can be written mathematically as

$$LR = \frac{P\left(E \mid H_p\right)}{P\left(E \mid H_d\right)},$$

where

- H_p is 'the suspect is the donor of the genetic data' (prosecutor's hypothesis),
- H_d is 'the suspect is unconnected to the crime' (defence attorney's hypothesis),

- $P(E \mid H_p)$ is often assumed to be 1 and
- $P(E \mid H_d)$ is the match probability.

The match probability is the probability that the suspect matches the haplotype found at the crime scene given that the suspect is unconnected to the crime, often translated to how probable it is that some random man's haplotype matches the haplotype found at the crime scene.

Methods to estimate the match probability are well established for traditional DNA profiles (based on autosomal STRs), see e.g. Balding and Nichols (1994), with most of them assuming statistical independence between the markers included in the profile. Due to the lack of recombination and, therefore, lack of statistical independence, the calculation of match probabilities is more challenging for lineage than for autosomal markers as described e.g. by Buckleton *et al.* (2011); Andersen *et al.* (2013a,c). In particular, when considering Y-STR haplotypes comprising up to 17 loci as Willuweit and Roewer (2009), the proportion of cases involving singletons, defined as haplotypes observed only once in a reference database augmented by the suspect profile, may become so large that use of traditional count estimates of the corresponding match probabilities becomes unsatisfactory. Therefore, better methods for modelling the haplotype distribution are needed such that satisfactory match probabilities can be calculated.

To detail the inference problem arising with singleton haplotypes, assume that a reference database of size n is given, and that a trace and suspect carry a new haplotype not yet observed in the database. Initially, the count estimator $1/(n+1)$ was used to derive match probabilities in such cases. However, this estimator is rather conservative because it is limited from below by the inverse of the database size. This is also demonstrated in simulation studies in this thesis. Therefore, a more advanced method referred to as 'haplotype surveying' was proposed (Roewer *et al.*, 2000; Krawczak, 2001) that tried to exploit the information about evolutionary relatedness inherent in a given database of Y-STR haplotypes. In view of the criticisms raised against it (Andersen, 2010; Brenner, 2010), the surveying method was later refined by Willuweit *et al.* (2011) and a new version is now implemented at the YHRD website (Roewer *et al.*, 2001; Willuweit and Roewer, 2009) (see `http://www.yhrd.org`). Brenner (2010) suggested an alternative, comparatively simple method of estimating the match probability for singletons for any kind of markers, the so-called 'κ correction' of the count estimator inspired by Robbins (1968). In short, the κ correction entails estimating a match probability by $(1-\kappa)/(n+1)$, where $\kappa = \alpha/(n+1)$ and α denotes the total number of singletons in the database.

In this part, two methods for calculating match probabilities are presented and compared to existing estimators like the κ correction by Brenner (2010).

Paper IV. *'Estimating trace-suspect match probabilities for singleton Y-STR haplotypes using coalescent theory'*

This chapter is based on Andersen *et al.* (2013a).

In this paper, a theory called coalescent theory was used to estimate haplotype frequencies. Coalescent theory tries to infer the genealogy or gene history of a population by using a database of haplotypes from the population. The analysis is done by assuming that the individuals in the population have a most recent common ancestor (MRCA) that can be inferred.

In principle, the genealogy that is most likely could be used, but because so many almost equally likely genealogies exist, a huge sample of these are often taken instead and each is weighted by its probability of occurring.

As might be speculated from the description, the method is rather computational intensive, which is why only relatively small datasets were analysed with this method in this study.

The method was implemented by modifying existing software (BATWING by Wilson *et al.* (2003)). Later, the method was implemented as an `R` (R Development Core Team, 2013) package called `rforensicbatwing` (Andersen and Wilson, 2013) (freely available open source software).

Paper V. *'Efficient forward simulation of Fisher-Wright populations with stochastic population size and neutral single step mutations'*

This chapter is based on Andersen and Eriksen (2012a).

When developing a statistical model, model control is of great importance. A model for the distribution of Y-STR haplotypes can be tested on real databases and compared to the results of other models, but the true population frequency of a haplotype is unknown. Hence, it is difficult to identify the errors.

One way to circumvent this problem is to simulate an entire population. Then, all the frequencies of all haplotypes are known. From this population, databases can be drawn and used by the models to estimate haplotype frequencies. These estimated frequencies can then be compared to the known ones such that the size of the errors can be estimated.

In this paper, a well known population model, the Fisher-Wright model of evolution by Fisher (1922, 1930, 1958); Wright (1931) with a single step mutation process by Ohta and Kimura (1973), was reformulated to facilitate computationally efficient simulations of even large populations. The method was implemented as an `R` (R Development Core Team, 2013) package called `fwsim` (Andersen and Eriksen, 2012b) (freely available open source software).

Paper VI. *'The discrete Laplace exponential family and estimation of Y-STR haplotype frequencies'*

This chapter is based on Andersen *et al.* (2013c).

An exponential family is a class of probability distributions that is well understood in probability theory such that inference can easily be made.

In this paper, an exponential family called the 'discrete Laplace distribution' was described. Its simple usage was exemplified by showing that it approximates a more complicated distribution by Caliebe *et al.* (2010) that arises in the Fisher-Wright model of evolution (Fisher, 1922, 1930, 1958; Wright, 1931) with a single step mutation process (Ohta and Kimura, 1973).

The theory for making inference in a mixture of multivariate, marginally independent, discrete Laplace distributions was then described. The model was used for estimating haplotype frequencies with lower prediction errors than those of other existing estimators like that of Brenner (2010).

Due to the known properties of exponential families, the calculations could be implemented and performed on a normal computer.

The method was implemented as an `R` (R Development Core Team, 2013) package called `disclapmix` (Andersen and Eriksen, 2013) (freely available open source software).

Paper VII. *'A gentle introduction to the discrete Laplace method for estimating Y-STR haplotype frequencies'*

This chapter is based on Andersen *et al.* (2013b).

Following the 'gentle introduction' genre for software, this paper is a gentle introduction to the discrete Laplace method. The method was described in a less technical manner than the original paper and the use of the software was demonstrated.

Paper VIII. *'Cluster analysis of European Y-chromosomal STR haplotypes using discrete Laplace distributions'*

This chapter is based on a preprint that has been submitted to Forensic Science International: Genetics (2013).

As already mentioned above in this introduction, lineage DNA markers can be used for population genetic analyses. Roewer *et al.* (2005) demonstrated how Y-STR markers could be used to infer recent historical events in the European Y-STR haplotype distribution. In this paper, using a completely different cluster analysis based on the discrete Laplace method that could be performed on a normal computer, we obtained similar results.

Because the discrete Laplace method is a probability model, other analyses than those similar to those of Roewer *et al.* (2005) were possible. For example, pairwise distances (between geographically separated samples) were also compared with those obtained using the AMOVA method by Excoffier *et al.* (1992) and good agreement was found. Furthermore, we investigated the homogeneity (uniformity of individuals in a population) in two different ways and found that the Y-STR haplotypes from e.g. Finland were relatively homogeneous as opposed to the relatively heterogeneous Y-STR haplotypes from e.g. Lublin, Eastern Poland and Berlin, Germany.

Paper IX. *'Efficient iteratively reweighted least squares for weighted two-way analysis of variance'*

This chapter is based on a paper that is in preparation for submission.

The implementation of the method described in paper VI used traditional inference techniques and worked well for moderately sized datasets (e.g. 13,000 haplotypes and 7 loci as analysed in paper VIII). For larger datasets as obtained from a yet unpublished collaborative YHRD study of 23 Y-STRs in various

populations (personal communication with Lutz Roewer and Michael Nothnagel) containing more than 18,000 haplotypes, the method can be greatly optimised by exploiting known model structure. In this paper, this optimisation is described in a slightly more general setup than actually needed for the method described in paper VI. This method was implemented in version 1.0 of the `R` package `disclapmix` and gives major speed-up compared to the original implementation.

4. Bibliography

Andersen, M. M. (2010) *Y-STR: Haplotype Frequency Estimation and Evidence Calculation*. Master's thesis, Aalborg University, Denmark. 4

Andersen, M. M., Caliebe, A., Jochens, A., Willuweit, S. and Krawczak, M. (2013a) Estimating trace-suspect match probabilities for singleton Y-STR haplotypes using coalescent theory. *Forensic Science International: Genetics*, **7**, 264–271. 4

Andersen, M. M. and Eriksen, P. S. (2012a) Efficient Forward Simulation of Fisher-Wright Populations with Stochastic Population Size and Neutral Single Step Mutations in Haplotypes. *Preprint, arXiv:1210.1773*. 5

Andersen, M. M. and Eriksen, P. S. (2012b) *fwsim: Fisher-Wright Population Simulation*. URL `http://CRAN.R-project.org/package=fwsim`. R package version 0.2-5. 5

Andersen, M. M. and Eriksen, P. S. (2013) *disclapmix: Discrete Laplace mixture inference using the EM algorithm*. URL `http://CRAN.R-project.org/package=disclapmix`. R package version 1.2. 6

Andersen, M. M., Eriksen, P. S. and Morling, N. (2013b) A gentle introduction to the discrete Laplace method for estimating Y-STR haplotype frequencies. *Preprint, arXiv:1304.2129*. 6

Andersen, M. M., Eriksen, P. S. and Morling, N. (2013c) The discrete Laplace exponential family and estimation of Y-STR haplotype frequencies. *Journal of Theoretical Biology*, **329**, 39–51. 4, 5

Andersen, M. M., Mogensen, H. S., Eriksen, P. S., Olofsson, J. K., Asplund, M. and Morling, N. (2013d) Estimating Y-STR allelic drop-out rates and adjusting for interlocus balances. *Forensic Science International: Genetics*, **7**, 327–336. 3

Andersen, M. M., Olofsson, J. K., Mogensen, H. S., Eriksen, P. S. and Morling, N. (2011) Estimating stutter rates for Y-STR alleles. *Forensic Science International: Genetics Supplement Series*, **3**, e192–e193. 2

Andersen, M. M. and Wilson, I. J. (2013) *rforensicbatwing: BATWING for calculating forensic trace-suspect match probabilities*. R package version 1.1. 5

Balding, D. J. and Nichols, R. A. (1994) DNA profile match probability calculation:

how to allow for population stratification, relatedness, database selection and single bands. *Forensic Sci. Int.*, **64**, 125–140. 4

Brenner, C. H. (2010) Fundamental problem of forensic mathematics – The evidential value of a rare haplotype. *Forensic Science International: Genetics*, **4**, 281–291. 4, 6

Buckleton, J., Krawczak, M. and Weir, B. (2011) The interpretation of lineage markers in forensic DNA testing. *Forensic Science International: Genetics*, **5**, 78–83. 4

Butler, J. M. (2001) *Forensic DNA Typing: Biology and Technology behind STR Markers*. Academic Press. 1, 2

Butler, J. M. (2005) *Forensic DNA Typing: Biology, Technology, and Genetics of STR Markers*. Academic Press, 2nd edn. 1, 2

Butler, J. M. (2010) *Fundamentals of Forensic DNA Typing*. Academic Press. 1, 2

Butler, J. M. (2012) *Advanced Topics in Forensic DNA Typing: Methodology*. Academic Press. 1, 2

Caliebe, A., Jochens, A., Krawczak, M. and Rösler, U. (2010) A Markov Chain Description of the Stepwise Mutation Model: Local and Global Behaviour of the Allele Process. *Journal of Theoretical Biology*, **266**, 336–342. 5

Cann, R., Stoneking, M. and Wilson, A. (1987) Mitochondrial DNA and human evolution. *Nature*, **325**, 31–36. 2

Evett, I. W. and Weir, B. S. (1998) *Interpreting DNA Evidence*. Sinauer Associates. 3

Excoffier, L., Smouse, P. E. and Quattro, J. M. (1992) Analysis of molecular variance inferred from metric distances among dna haplotypes: Application to human mitochondrial dna restriction data. *Genetics*, **131**, 479–491. 6

Fisher, R. A. (1922) On the Dominance Ratio. *Proc. Roy. Soc. Edin.*, **42**, 321–341. 5

Fisher, R. A. (1930) *The Genetical Theory of Natural Selection*. Oxford: Clarendon Press. 5

Fisher, R. A. (1958) *The Genetical Theory of Natural Selection*. New York: Dover, 2nd revised edn. 5

Gill, P., Jeffreys, A. J. and Werrett, D. J. (1985) Forensic application of DNA fingerprints. *Nature*, **318**, 577–579. 1

Krawczak, M. (2001) Forensic evaluation of Y-STR haplotype matches: a comment. *Forensic Science International*, **118**, 114–115. 4

Ohta, T. and Kimura, M. (1973) A Model of Mutation Appropriate to Estimate the Number of Electrophoretically Detectable Alleles in a Finite Population. *Genet. Res.*, **22**, 201–204. 5

Olofsson, J. K., Andersen, M. M., Mogensen, H. S., Eriksen, P. S. and Morling, N. (2012) Sequence variants of allele 22 and 23 of DYS635 causing different stutter rates. *Forensic Science International: Genetics*, **6**, e161–e162. Letter to Editor. 2

Piazza, A., Mattiuz, P. and Ceppellini, R. (1969) [Combination of haplotypes of the HL-A system as a possible mechanism for gametic or zygotic selection]. *Haematologica*, **54**, 703–720. Article in Italian. 1

R Development Core Team (2013) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org`. ISBN 3-900051-07-0. 5, 6

Robbins, H. E. (1968) Estimating the Total Probability of the Unobserved Outcomes of an Experiment. *The Annals of Mathematical Statistics*, **39**, 256–257. 4

Roewer, L. (2009) Y chromosome STR typing in crime casework. *Forensic Sci Med Pathol*, **5**, 77–84. 1

Roewer, L., Croucher, P. J. P., Willuweit, S., Lu, T. T., Kayser, M., Lessig, R., de Knijff, P., Jobling, M. A., Tyler-Smith, C. and Krawczak, M. (2005) Signature of recent historical events in the European Y-chromosomal STR haplotype distribution. *Human Genetics*, **116**, 279–291. 2, 6

Roewer, L., Kayser, M., de Knijff, P. *et al.* (2000) A new method for the evaluation of matches in non-recombining genomes: application to Y-chromosomal short tandem repeat (STR) haplotypes in European males. *Forensic Science International*, **114**, 31–43. 4

Roewer, L., Krawczak, M., Willuweit, S. *et al.* (2001) Online reference database of European Y-chromosomal short tandem repeat (STR) haplotypes. *Forensic Science International*, **2-3**, 106–113. 4

Sibille, I., Duverneuil, C. *et al.* (2002) Y-STR DNA amplification as biological evidence in sexually assaulted female victims with no cytological detection of spermatozoa. *Forensic Sci. Int.*, **125**, 212–216. 1

Sullivan, K., Hopgood, R., Lang, B. and Gill, P. (1991) Automated amplification and sequencing of human mitochondrial DNA. *Electrophoresis*, **12**, 17–21. 1

Willuweit, S., Caliebe, A., Andersen, M. M. and Roewer, L. (2011) Y-STR Frequency Surveying Method: A critical reappraisal. *Forensic Science International: Genetics*, **5**, 84–90. 4

Willuweit, S. and Roewer, L. (2009) Y chromosome haplotype reference database (YHRD): Update. *Forensic Science International: Genetics*, **1**, 83–87. 4

Wilson, I. J., Weale, M. E. and Balding, D. J. (2003) Inferences from DNA Data: Population Histories, Evolutionary Processes and Forensic Match Probabilities. *Journal of Royal Statistical Society Series A*, **166**, 155–201. 5

Wright, S. (1931) Evolution in Mendelian populations. *Genetics*, **16**, 97–159. 5

Part 1

Error modelling

Paper I & II

Estimating stutter rates of Y-STR alleles

Author list Mikkel Meyer Andersen, *Aalborg University, Denmark*
Jill Katharina Olofsson, *University of Copenhagen, Denmark*
Helle Smidt Mogensen, *University of Copenhagen, Denmark*
Poul Svante Eriksen, *Aalborg University, Denmark*
Niels Morling, *University of Copenhagen, Denmark*

Summary Stutter peaks are artefacts that arise during PCR amplification of short tandem repeats. Stutter peaks are especially important in forensic case work with DNA mixtures. The aim of the study was primarily to estimate the stutter rates of the AmpFlSTR Yfiler kit. We found that the stutter rates differ at the allelic level and with the parental peak height.

Publication info This chapter is a bringing together the following publications:

Andersen MM, Olofsson JK, Mogensen HS, Eriksen PS, Morling N (2011). *Estimating stutter rates for Y-STR alleles.* Forensic Science International: Genetics Supplement Series; **3**(1):e192-e193.

Olofsson JK, Andersen MM, Mogensen HS, Eriksen PS, Morling N (2012). *Sequence variants of allele 22 and 23 of DYS635 causing different stutter rates.* Forensic Science International: Genetics. Letter to editor; **6**(6):e161-e162.

1. Introduction

Stutter peaks are artefacts that arise during PCR amplification of short tandem repeats that are highly polymorphic genetic markers commonly used in forensic genetics (Butler, 2006). Stutter peaks are especially important in forensic case work when DNA mixtures are analysed. To analyse mixtures properly, good estimates of stutter rates – stutter peak height divided with the parental peak height (Brookes *et al.*, 2012) – must be available. The aim of the study was (1) to estimate the stutter rates of the AmpFlSTR Yfiler kit (Applied Biosystems – AB), (2) to investigate the stutter rates at the allelic level, and (3) to test if the stutter rate changed with the parental peak height.

2. Material and methods

Two 1.2 mm punches of FTA® cards (Whatman) with buccal samples from each of 360 persons were amplified in 10 μl reaction volume with AmpFlSTR® Yfiler® kit with 27 cycles. PCR products were separated on an AB3130xl (AB) and fragments analysed using GeneScan 3.7 and Genotyper 3.7 (AB) with 5 RFU threshold. For each sample, the highest peak at each locus was taken as the parental peak if the height was between 50 and 7,000 RFU. The heights of the parental and -1 repeat stutter peaks were further analysed.

2.1. *Simple linear regression*

The data was first analysed using weighted linear regression for each locus with stutter peak height as the response variable and parental peak height as the explanatory variable. The inverse parental peak height was used as weight to incorporate that the variance increases with the signal strength. The model included an intercept to reflect the fact that the stutter rate – defined as the stutter height divided by the parental peak height – changed with the height of the parental peak height.

This model can be written as

$$\text{StutterHeight} = \beta_0 + \beta_1 \cdot \text{ParentHeight}.$$

Note, that with this model, the stutter rate has the form

$$\frac{\text{StutterHeight}}{\text{ParentHeight}} = \frac{\beta_0}{\text{ParentHeight}} + \beta_1.$$

This results in the following interpretation when assuming $\beta_1 > 0$ and ParentHeight > 1: For a positive intercept (β_0), the stutter rate decreases when ParentHeight increases.

2.2. *Multiple linear regression*

Later the data was analysed using a weighted multiple linear regression for each locus with stutter peak height as the response variable and parental peak height

together with allele lengths and their interaction as the explanatory variables, namely

$$\text{StutterHeight} = \beta_0 + \beta_0 \cdot \text{Allele} + \beta_2 \cdot \text{ParentHeight} + \beta_3 \cdot \text{Allele} \cdot \text{ParentHeight}.$$

Note, that with this model, the stutter rate has the form

$$\begin{aligned}\frac{\text{StutterHeight}}{\text{ParentHeight}} &= \frac{\beta_0}{\text{ParentHeight}} + \beta_1 \frac{\text{Allele}}{\text{ParentHeight}} + \beta_2 + \beta_3 \cdot \text{Allele} \\ &= \text{ParentHeight}^{-1} \cdot (\beta_0 + \beta_1 \cdot \text{Allele}) + \beta_2 + \beta_3 \cdot \text{Allele}.\end{aligned}$$

Results obtained with the multiple regression model was compared to those obtained with the Kazam stutter rates supplied by AB.

3. Results and discussion

For the weighted multiple linear regression, the adjusted R^2 values varied between 82.5% (DYS438) and 98.9% (DYS390). Besides DYS438, only two additional loci had an adjusted R^2 value below 90% (DYS635 had an adjusted R^2 value of 85.8% and DYS448 had an adjusted R^2 value of 89.7%).

In Figure 1, a simple linear regression for DYS390 is shown.

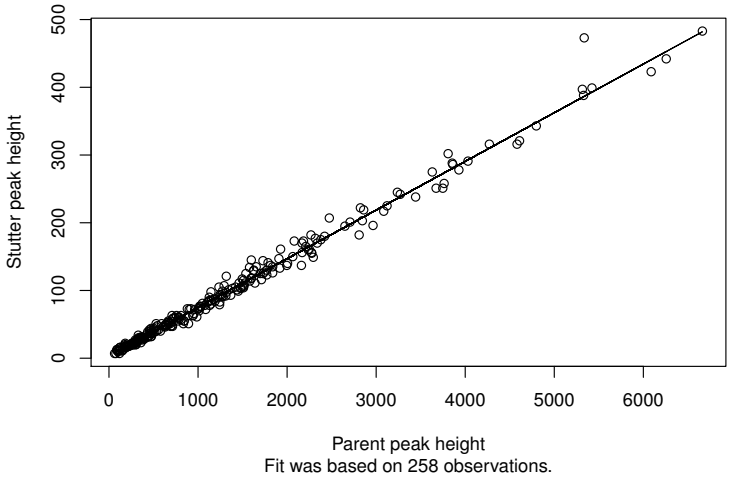


Figure 1. Stutter peak heights for DYS390 allele 23. As seen, the linear dependence is large.

DYS635 yielded a poor fit, especially for allele 23. This was investigated by looking at the simple linear regression shown in Figure 2, where two groups of stutter rates were identified. Sequencing of 14 samples using BigDye Termination v1.1 Cycle Sequencing Kit showed that 9 samples had sequences with longest uninterrupted stretch (LUS) of 9 repetitive units and 5 samples

had sequences with LUS equal to 13 repetitive units. This discrepancy is due to a complex structure with several repetitive sequences of varying length together with intervening sequences (as defined by Urquhart *et al.* (1994)). The sequence variants were in accordance with the previously published sequences of DYS635 (Gusmao *et al.*, 2002). All samples with LUS 13 were in the group with high stutter rates and all samples with LUS 9 were in the group with lower stutter rates.

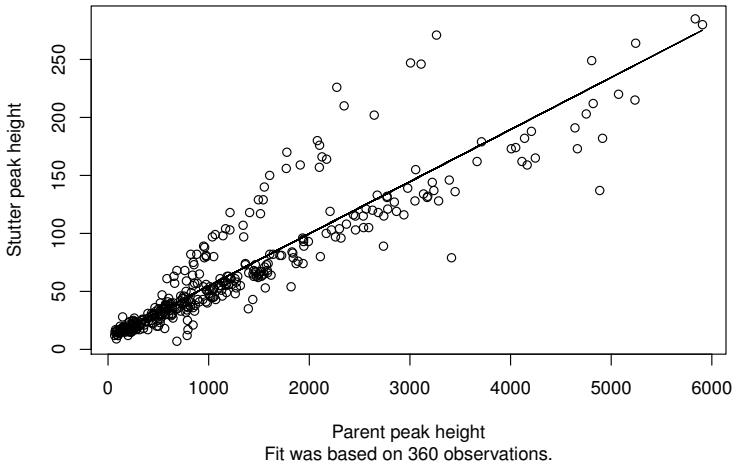


Figure 2. Stutter peak heights for DYS635 allele 23. As seen, there are two groupings. Sequencing of 14 samples revealed that the grouping is due to differing longest uninterrupted stretch (LUS) between the groups. All 5 samples with LUS 13 were in the group with high stutter rates and all 9 samples with LUS 9 were in the group with lower stutter rates.

DYS438 and DYS448 seem to fit poorly due to a greater spread of the stutter rates given the parental peak heights. The relationships are still linear. The reason for this is not known; DYS438 has a simple structure (as opposed to e.g. DYS635), whereas DYS448 has a more complex structure although with no LUS variants like DYS635.

Table 1 compares the predicted stutter heights for DYS389I allele 12-14 using the weighted multiple linear regression model with the Kazam constant stutter rate for various parental peak heights. The Kazam stutter rates are upper bounds whereas the estimates given in the present paper are means. These are two conceptually different approaches. We have taken the approach of using the mean because an upper bound is not consistently conservative, it depends on the situation.

	Parental peak height					
	50 RFU		500 RFU		2000 RFU	
	Stutter height/rate		Stutter height/rate		Stutter height/rate	
Allele	RFU	%	RFU	%	RFU	%
12	2.8	5.5 %	23.6	4.7 %	93.2	4.7 %
13	4.5	9.0 %	29.7	5.9 %	113.8	5.7 %
14	6.2	12.5 %	35.8	7.2 %	134.3	6.7 %
Kazam	5.9	11.79 %	59.0	11.79 %	235.8	11.79 %

Table 1. Stutter height and rate predictions for DYS389I by allele and parental heights using weighted multiple linear regression. Kazam refers to Applied Biosystems' recommended stutter filter.

4. Conclusion

Stutter rates differ on the allelic level, hence one stutter rate per locus is not optimal. Stutter rates seem to increase with the numbers of Y-STR repeats as seen in Table 1. Applied Biosystems' recommended stutter filter rates seem to be too high in general, which can cause problems in analysing DNA mixtures. Table 1 also show, remembering that stutter rate is stutter peak height divided by parental peak height, that intercepts need to be included in the model because the stutter rate actually does change with the parental peak height.

The constructed weighted multiple linear regression models seem to predict stutter heights quite well on almost all loci using allele and parental peak height as explanatory variables. This gives an easy way of predicting stutter heights. Intra-allelic problems exist, especially among DYS635 alleles that have a complex structure with several repetitive sequences of varying lengths together with intervening sequences that causes different stutter rates among alleles with the same length; this complicates mixture analysis greatly.

5. Bibliography

Brookes, C., Bright, J., Harbison, S. and Buckleton, J. (2012) Characterising stutter in forensic STR multiplexes. *Forensic Science International: Genetics*, **6**, 58–63. 14

Butler, J. M. (2006) Genetics and genomics of core short tandem repeat loci used in human identity testing. *Journal of Forensic Science*, **51**, 253–265. 14

Gusmao, L., Gonzalez-Neira, A., Alves, C., Lareu, M., Costa, S., Amorim, A. and Carracedo, A. (2002) Chimpanzee homologous of human Y specific STRs – A comparative study and a proposal for nomenclature. *Forensic Science International*, **126**, 129–136. 16

Urquhart, A., Kimpton, C. P., Downes, T. J. and Gill, P. (1994) Variation in short tandem repeat sequences - a survey of twelve microsatellite loci for use as forensic identification markers. *Int. J. Med.*, **107**, 13–20. 16

Paper III

Estimating Y-STR allelic drop-out rates and adjusting for interlocus balances

Author list

Mikkel Meyer Andersen, *Aalborg University, Denmark*
Helle Smidt Mogensen, *University of Copenhagen, Denmark*
Poul Svante Eriksen, *Aalborg University, Denmark*
Jill Katharina Olofsson, *University of Copenhagen, Denmark*
Maria Asplund, *University of Copenhagen, Denmark*
Niels Morling, *University of Copenhagen, Denmark*

Summary

Y chromosome short tandem repeats (Y-STRs) are valuable genetic markers in certain areas of forensic case-work. However, when the Y-STR DNA profile is weak, the observed Y-STR profile may not be complete – i.e. locus drop-out may have occurred. Another explanation could be that the stain DNA did not have a Y-STR allele that was detectable with the method used (the allele is a 'null allele'). If the Y-STR profile of a stain is strong, one would be reluctant to consider drop-out as a reasonable explanation of lack of a Y-STR allele and would maybe consider 'null allele' as an explanation. On the other hand, if the signal strengths are weak, one would most likely accept drop-out as a possible explanation. We created a logistic regression model to estimate the probability of allele drop-out with the Life Technologies/Applied Biosystems AmpF*l*STR® Yfiler® kit such that the trade-off between drop-outs and null alleles could be quantified using a statistical model. The model to estimate the probability of drop-out uses information about locus imbalances, signal strength, the number of PCR cycles, and the fragment size of Yfiler. We made two temporarily separated experiments and found no evidence of temporal variation in the probability of drop-out. Using our model, we found that for 30 PCR cycles with a 150 bp allele, the probability of drop-out was 1:5,000 corresponding to the average estimate of the probability of Y-STR null alleles at a signal strength of 1,249 RFU. This means that the probability of a null allele is higher than that of an allele drop-out at e.g. 4,000 RFU and the probability of drop-out is higher than that of a null allele at e.g. 75 RFU.

Publication info

This paper was published as:

> Andersen MM, Mogensen HS, Eriksen PS, Olofsson JK, Asplund M, Morling N (2013). *Estimating Y-STR allelic drop-out rates and adjusting for interlocus balances.* Forensic Science International: Genetics; **7**(3):327-336.

The version here is the journal version with minor typographical corrections.

1. Introduction

Y chromosome short tandem repeats (Y-STRs) are valuable genetic markers in forensic case-work, especially in sexual assault cases where only small amounts of DNA from a male perpetrator is found in combination with a large amount of DNA from a female victim (Gill *et al.*, 2001; Gusmao *et al.*, 2006; Roewer, 2009). The reason for this is that the routine investigation of autosomal STRs, in such cases, will result in a DNA profile of the female victim, while investigations of Y-chromosome markers will result in a male Y-STR profile even if the amount of female DNA is more that 1,000 times larger than that of male DNA (Prinz *et al.*, 1997). The weight of the evidence of matching Y-STR DNA profiles from e.g. a scene of crime and a suspect may be estimated by likelihood principles (Morling *et al.*, 2002; Gill *et al.*, 2006). The weight of the evidence is usually presented as a likelihood ratio (LR) of

$$\frac{Pr(\text{Y-STR profile} \mid \text{the DNA comes from the suspect})}{Pr\left(\text{Y-STR profile} \,\middle|\, \begin{array}{l}\text{the DNA comes from a random}\\ \text{person not related to the suspect}\end{array}\right)}.$$

To be able to calculate this, one must have a sound estimate of the probability of observing the Y-STR profile among random individuals in the relevant population. This is a problem in itself (Roewer *et al.*, 2000; Krawczak, 2001; Brenner, 2010; Buckleton *et al.*, 2011; Andersen *et al.*, 2013). The other part of the LR is the probability of the Y-STR profile under the assumption that it comes from the suspect. This is easy if the Y-STR profiles of the crime scene sample and the suspect are identical – the probability is 1. However, when the amount of Y-STR DNA is small and the Y-STR DNA profile is weak, the observed Y-STR profile may not be complete – i.e. locus drop-out may have occurred. This phenomenon is often considered of minor importance, and the lack of result from a locus is often ignored under the assumption that the phenomenon was due to locus drop-out. However, another explanation could be that the stain DNA did not have a Y-STR allele that was detectable with the method used – typically due to a SNP in the primer binding regions of around the Y-STR (Butler, 2005; Budowle *et al.*, 2008). The average frequency of such 'null alleles' is approximately 1:5,000 = 0.02% (in release 39 of `http://www.yhrd.org` (Roewer *et al.*, 2001; Willuweit and Roewer, 2009) there were 219 null alleles amount 1,111,984 alleles in total). If the Y-STR profile of a stain is strong with signal strength of e.g. 4,000 RFU on an AB3130xl, drop-out is highly unlikely (Tvedebrink *et al.*, 2009, own unpublished observations). However, if the signal strength is e.g. 75 RFU, the probability of drop-out is approximately 20% (cf. Figure 10), and drop-out must be included as a possible explanation.

Although the risk of drop-out may not seem so important for Y-STRs as for autosomal STRs (Tvedebrink *et al.*, 2009, 2011a), it should still be considered. We have investigated the drop-out risk of the AmpF*l*STR® Yfiler® (Life Technologies/Applied Biosystems) when using the kit with 28, 29, and 30 PCR cycles. We offer an easy method based on logistic regression analysis to estimate the drop-out risk of Y-STRs.

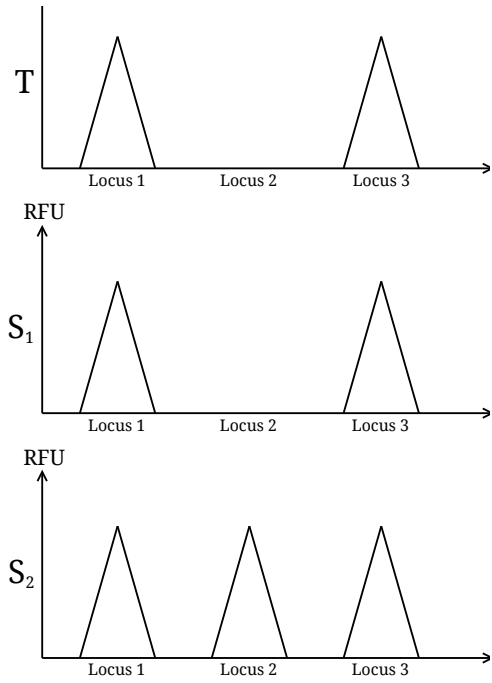


Figure 1. An example that motivates to estimate the probability of allele drop-out. Assume that the topmost electropherogram (EPG) denoted by 'T' was obtained from the evidence found at the crime scene and the two ones below are from two reference samples, 'S_1' and 'S_2'. Now, which reference sample is most consistent with 'T'? 'S_1' can explain 'T' by a null allele and 'S_2' can explain 'T' by an allele drop-out. If the peaks in 'T' are around e.g. 75 RFU, then we might suspect allele drop-out that would make 'S_2' consistent with 'T'. On the other hand, if the peaks in 'T' are around e.g. 4,000 RFU, we would not suspect an allele drop-out, but instead suspect a null allele. Thus, in order to make a better analysis, we need a model to estimate the probability of allele drop-out compared to that of a null allele.

1.1. *Motivating example*

A simple example that motivates the evolution of the probability of allele drop-out is given in Figure 1. A more complicated example is as follows: For the sake of argument, assume that the probability of a null allele at a locus is 1 : 5,000 = 0.02% (which correspond to the number of null alleles in release 39 of `http://www.yhrd.org` (Roewer *et al.*, 2001; Willuweit and Roewer, 2009)). Assume a two person mixture, where all but one locus has two peaks, each of height 4,000 RFU. The last locus only has one peak of height 4,000 RFU. The profile is well-balanced and there is no evidence of two shared alleles at this locus as this in theory would result in a peak of 8,000 RFU. At 4,000 RFU, the probability of drop-out is approximately 1:100,000 (cf. Table 1). This should be compared to the probability of a null allele (1:5,000), which gives odds of 20 for a null allele compared to a drop-out.

Now, assume a two person mixture where all but one loci have two peaks, each of height 75 RFU. The last locus only has one peak of height 75 RFU. Again,

we have a well-balanced profile where there is no evidence of two shared alleles at this locus. At 75 RFU, the probability of drop-out is approximately 1:5 (cf. Figure 10). This should be compared to the probability of a null allele (1:5,000), which gives odds of 1,000 for a drop-out compared to a null allele.

2. Materials and methods

2.1. *Experiments*

Two sets of controlled experiments were conducted at The Section of Forensic Genetics, Department of Forensic Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Denmark. For estimating the drop-out probability, eight different male DNA samples were diluted into 14 different concentrations and amplified in triplicates at 28, 29 and 30 thermocycles using the AmpF*l*STR® Yfiler® (Life Technologies/Applied Biosystems) amplification kit. The first set of experiments were conducted with DNA from four males. In the second set of experiments, DNA from four other males was investigated. In the first experiment, only data from 28 and 30 thermocycles were available.

For dilution series, blood samples were taken from eight males. Genomic DNA was extracted with the EZ1 Investigator kit (Qiagen) using a BioRobot EZ1 (Qiagen) or with PrepFiler™ Express Forensic DNA Extraction Kit (AB) using an Automate Express™ robot (AB). Each DNA sample was quantified in triplicate using the Quantifiler® Y Human Male DNA Quantification Kit (AB) with Human Genomic DNA Male (G147A, Promega) as the quantification standard on an ABIPrism 7000 (AB) or an ABIPrism 7500 (AB). The median DNA concentration was used. Each sample was diluted with water to DNA concentrations of 100 pg/μl or 1,000 pg/μl. Dilution series were performed with serial dilutions to give 14 different DNA concentrations in the range 0.75-150 pg/μl.

A total of 5 or 10 μl of the diluted samples was added to the PCR mixture and each sample was amplified in triplicate with the AmpF*l*STR® Yfiler® PCR Amplification Kit (AB) as recommended by the manufacturer in an 96-Well GeneAmp® PCR System 9700 (AB) amplifying with 28, 29 and 30 thermocycles. The resulting amount of DNA in the PCR reactions ranged from 7.5-1,000 pg.

One μl of the amplificate together with 15 μl HiDi Formamide (AB) was analysed on an ABI Prism 3130xl Genetic Analyzer (AB) using POP4 (AB) as the polymer and 3 kV injection voltage for 10 seconds. DNA fragments were detected, fragment sizes were estimated, and alleles were assigned using GeneMapper 3.2 (AB) or GeneScan 3.7 with GenoTyper 3.7 (both AB) with a detection threshold of 15 RFU and no filter applied. A detection threshold of 50 RFU was used, which is also the detection threshold for drop-out. Peaks between 15 RFU and 50 RFU were included for improving statistical modelling.

The DNA profiles included only one allele per locus except for the DYS385a/b locus. Seven profiles had two alleles, and a single profile had one allele at the DYS385a/b locus.

The protocols were approved by the Danish ethical committee (KF-01-037/93

and H-1-2011-081).

2.2. *Data*

All data analysis was performed using the statistical software R (R Development Core Team, 2013).

In Figure 2, the proportion of dropped out Y-STR loci given the expected DNA concentration and the number of PCR cycles for the sample is shown. In Figure 3, the experiment is also included as a dependent variable.

No drop-out occured when the expected DNA concentration was greater than 100 pg/μl, which is why concentrations higher than 100 pg/μl are not shown in the Figure 2 and Figure 3.

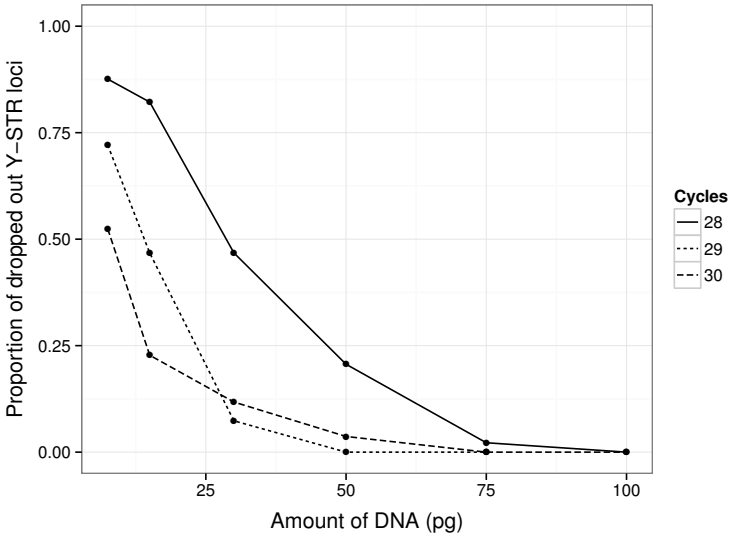


Figure 2. The proportion of dropped out Y-STR loci depending on the amount of DNA. No drop-out occurred when the amount of DNA was greater than 100 pg.

2.3. *Estimating interlocus balances*

The AmpF*l*STR Yfiler amplification kit is not well balanced between loci, which is depicted in Figure 4. This means that locus balances need to be considered in the drop-out model. In this section, a model for estimating interlocus balances is described.

Due to the lack of accuracy and reproducibility in quantification, we could not use the quantified DNA amount in the model of the signal strength. Instead, we introduced an individual signal strength for each sample denoted by S_i for samples $i = 1,2,\ldots,n$. The signal strength can be described as the mean peak height weighted by the interlocus balances. We will now discuss the modelling of this in detail.

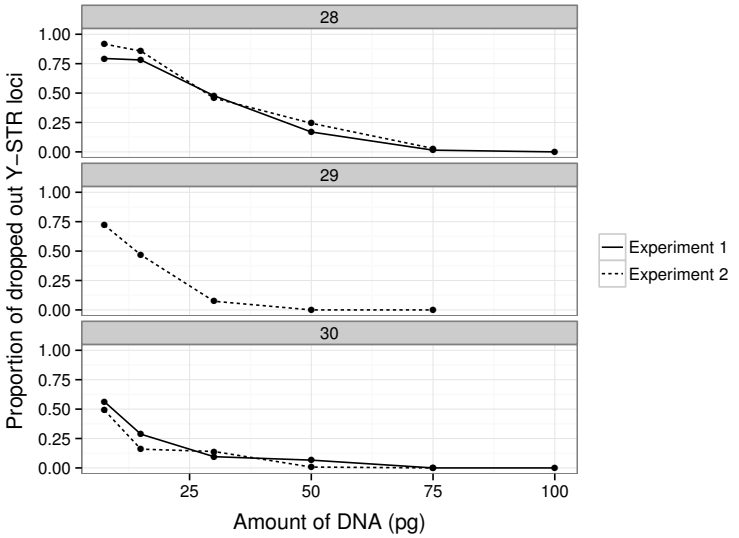


Figure 3. The proportion of dropped out Y-STR loci given the amount of DNA, cycles and experiment. No drop-out occurred when the amount of DNA was greater than 100 pg.

Let x_{ij} be the peak height at the j^{th} locus for the i^{th} sample for $j = 1,2,\ldots,r$ and $i = 1,2,\ldots,n$, where r is the number of loci and n is the number of samples. Then, we assume that $\log x_{ij}$ is normally distributed with a mean value depending on the sample and locus. In a statistical notation, where $N(\mu,\sigma^2)$ denotes a normal distribution with a mean value μ and the variance σ^2, we assume that

$$\log x_{ij} \sim N\left(\theta_j + \log S_i, \sigma^2\right), \tag{1}$$

where θ_j is the locus balance for the j^{th} locus and S_i is the signal strength for the i^{th} sample.

We impose constraints on the θ_j's such that

$$\sum_{j=1}^{r} \theta_j = 0.$$

As the linear model stated in Equation (1) assuming Equation (2) is a linear regression model, we checked it on samples with full profiles (samples with no drop-out) using the linear model fit function `lm` in the statistical software `R` (R Development Core Team, 2013). The adjusted R^2 value was 93.7% with both locus and sample as statistically significant factors. The resulting interlocus balances, θ_j, are depicted in Figure 5.

For locus DYS385a/b, only one locus balance is estimated based on the sum of the peak heights of 2 alleles (7 profiles) and the peak height for 1 allele (1 profile). Later, for signal strength estimation, DYS385 was treated as two loci, 'DYS385a' and 'DYS385b', each with locus balance $\theta' = \theta/2$, where θ is this estimated locus balance for the sum of the DYS385a/b peak heights.

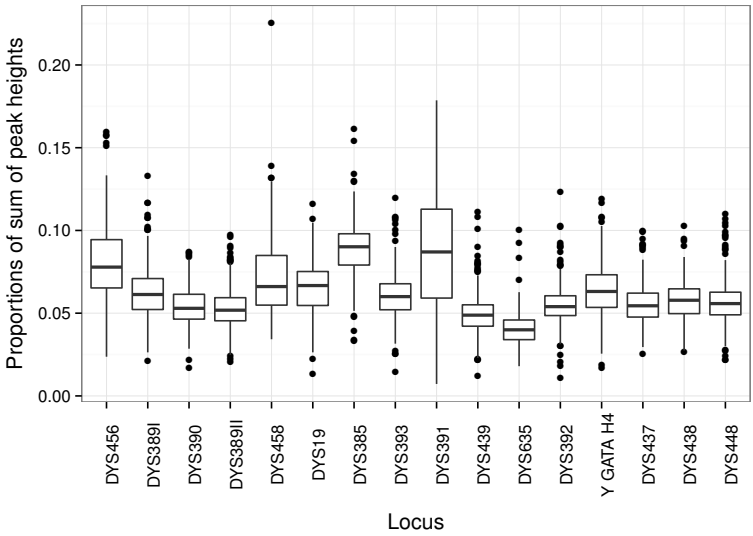


Figure 4. Interlocus balances of the peak heights at the Y-STR loci. To explain the box-and-whiskers plot, let q_p be the p% quantile. The box contains the middle 50% of the observations (from the 25% quantile, q_{25}, to the 75% quantile, q_{75}). The horizontal line in the box displays the median (50% quantile, q_{50}). The end of the lower whisker is the lowest datapoint greater than $q_{25} - 1.5 \times IQR$, where IQR is the interquartile range given by $q_{75} - q_{25}$ (the height of the box). The end of the upper whisker is the greatest datapoint lower than $q_{75} + 1.5 \times IQR$. The points are outliers that are either lower than $q_{25} - 1.5 \times IQR$ or greater than $q_{75} + 1.5 \times IQR$.

2.4. *Estimating signal strength*

Other studies on drop-outs, e.g. those of Tvedebrink *et al.* (2009, 2011a), use the signal strength as a predictor of the drop-out probability. We investigate the same predictor here. Due to the lack of balance of the Yfiler kit as described in Section 2.3, the signal strength must be modelled somewhat differently. Another difference in the modelling is that we incorporate the knowledge that some of the peaks may have dropped out by using a truncated probability distribution.

When we estimated interlocus balances on full profiles, we used the model in Equation (1), Section 2.3. Now, when we have drop-outs, a slightly different model for the peak heights was used instead, namely

$$\log x_{ij} \sim N_{\log t}\left(\theta_j + \log S_i, \sigma_i^2\right), \tag{2}$$

where $N_{\log t}(\cdot,\cdot)$ denotes a normal distribution truncated below $\log t$ (meaning that there is no observation less than $\log t$, where t is known and we have information about the number of observations being truncated). In forensic genetics, t is the detection threshold. Often the value $t = 50$ RFU is used, which we also used. As before, x_{ij} is the peak height at the j^{th} locus and the i^{th} sample, θ_j is the locus balance for the j^{th} locus and S_i is the signal strength for the i^{th} sample.

Now, assume that the interlocus balances estimated using Equation (1) are known. This is a reasonable assumption and it makes inference about the signal strength, S_i, easier.

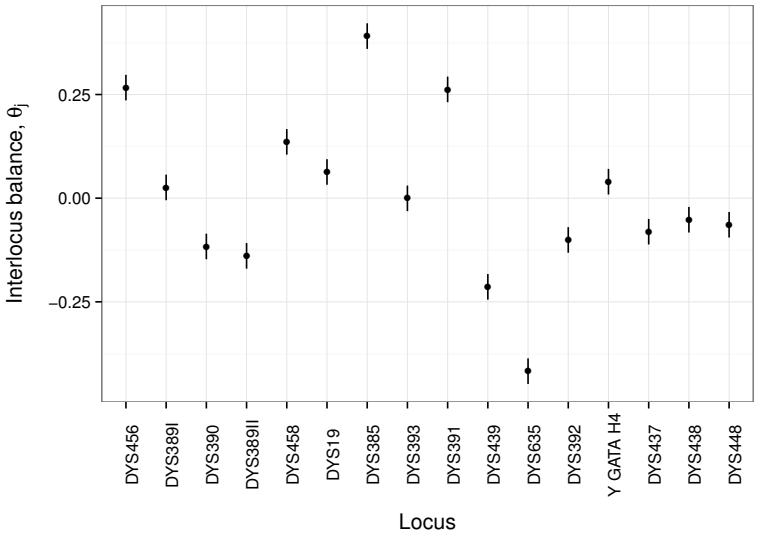


Figure 5. Interlocus balances, θ_j, from the model $\log x_{ij} \sim N(\theta_j + \log S_i, \sigma^2)$ with 95% confidence intervals. Note, that all interlocus balance estimates have the same variance due to the balanced design (all samples are full profiles).

The goal is to estimate S_i and use it as a proxy for the signal strength by using the peaks above 50 RFU, their heights and implicitly peaks that have dropped out.

If we assume that the interlocus balances, θ_j, are known, then the model for one sample is

$$\log x_j \sim N_{\log t}\left(\theta_j + \log S, \sigma^2\right). \tag{3}$$

Let $J \subseteq \{1,2,\ldots,r\}$ denote the set of loci that did not drop out and $J^C = \{1,2,\ldots,r\} \setminus J$, where $\setminus$ means set difference, the set of loci that dropped out. The likelihood of the model in Equation (3) for one sample $\{x_j\}_{j\in J}$ is then given by

$$\begin{aligned} L\left(\log S, \sigma^2; \{x_j\}_{j\in J}\right) &= \prod_{j=1}^{r} L_j \\ &= \prod_{j\in J^C} \Phi\left(\frac{\log t - (\theta_j + \log S)}{\sigma}\right) \times \prod_{j\in J} \sigma^{-1}\phi\left(\frac{\log x_j - (\theta_j + \log S)}{\sigma}\right), \end{aligned} \tag{4}$$

where L_j is the likelihood contribution from the j^{th} locus, Φ is the cumulative distribution function for the standard normal distribution and ϕ is the probability density function of the standard normal distribution. The first product sign, $\prod_{j\in J^C}$, collects the likelihood contribution of the loci that dropped out because $\Phi\left(\frac{\log t - (\theta_j + \log S)}{\sigma}\right)$ is the probability of observing a value less than $\log t$ in a $N\left(\theta_j + \log S, \sigma^2\right)$ distribution. The second product sign, $\prod_{j\in J}$, collects the likelihood

contribution from the loci that did not drop out because $\sigma^{-1}\phi\left(\frac{\log x_j-(\theta_j+\log S)}{\sigma}\right)$ is the probability of observing the value $\log x_j$ in a $N\left(\theta_j+\log S,\sigma^2\right)$ distribution.

For a sample $\{x_j\}_{j\in J}$, the likelihood in Equation (4) can be optimised numerically using the `optim` functionality in `R` (R Development Core Team, 2013) to obtain the estimate $\log\hat{S}$. Note, that if we have a full profile, ($J^C=\emptyset$), then the optimum of Equation (4) is $\log\hat{S}=r^{-1}\sum_{j=1}^{r}(\log x_j-\theta_j)=r^{-1}\sum_{j=1}^{r}\log x_j$. In other words, for a full profile, the log of the signal strength is the average of the log peak heights because the sum of the locus balances is 0. Also, note that at least two loci are required because both $\log S$ and σ^2 must be estimated.

If the information about truncation is ignored, then the crude estimator

$$\log\hat{S}_{\text{crude}}=\frac{1}{r-k}\sum_{j\in J}(\log x_j-\theta_j) \tag{5}$$

can be used, where $k=\left|J^C\right|$ is the number of loci dropped out. The crude estimator is expected to be greater than the likelihood estimator because it does not incorporate knowledge of the loci dropped-out and the estimate is decreased because the peaks dropped-out are known to be smaller than 50 RFU. In Figure 6, the signal strength estimator based on optimising the likelihood in Equation (4) is compared to the crude estimator in 5 using all the data from profiles with at least two loci not dropped out. This figure shows that the crude estimator in Equation (5) is greater than the likelihood based estimator in Equation (4).

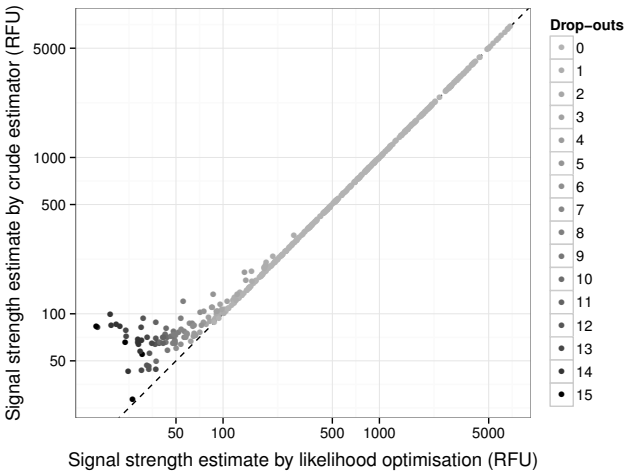


Figure 6. Comparison of the signal strength estimator based on the likelihood Equation (4) and the crude estimator Equation (5) based on all data with at least two loci not dropped out. The line has slope 1 and intercept 0 corresponding to a 1:1 correlation. The crude estimator is expected to be greater than the likelihood estimator (see the text for the arguments), which is supported by this figure (because the points are above the line).

Estimators of truncated normal distributions are treated by Persson and Rootzen (1977), but locus imbalances make things complicated, which is why we use the numerical optimisation.

Optimising Equation (4) makes it possible to estimate the signal strengths, $\hat{S}$, for all samples with at least two loci not dropped out. Only these samples with at least two loci not dropped out are used. In principle, the crude estimator Equation (5) could be used, but as described previously and shown in Figure 6, this would result in too large signal strengths for samples with only one locus. Another option would be to estimate the overall variance σ^2 such that only one observation would be needed to estimate the one parameter S. As shown in Figure 7, the variance for low signal strengths is probably too large to obtain a reasonable overall estimate.

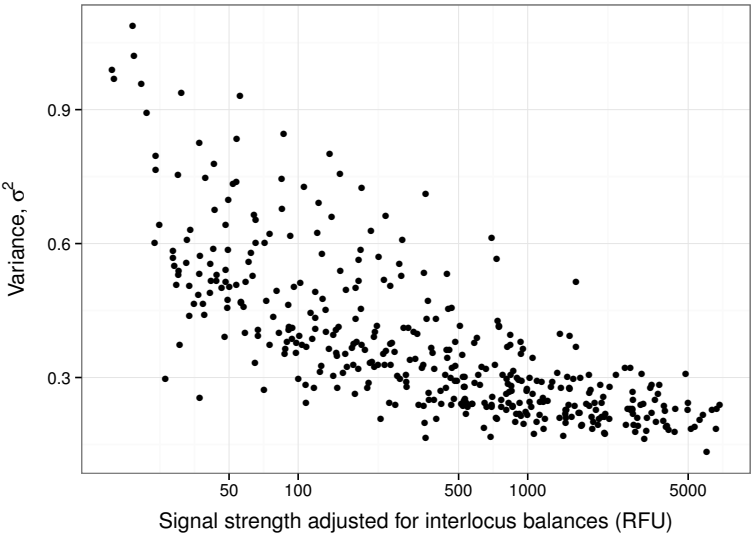


Figure 7. The variance, σ_i^2, for each sample given the signal strength, S_i, based on optimising Equation (4). The variance, σ_i^2, decreased with the signal strength.

In Figure 8, the correlation between the DNA concentration and the signal strength given the number of PCR cycles is shown. In Figure 9, the correlation between signal strength and the propotion of loci dropped out is depicted.

2.5. *Modelling drop-out probability*

As done in other studies, e.g. those of Tvedebrink *et al.* (2009, 2011a), logistic regression (Hosmer and Lemeshow, 2000; Agresti, 2002) of the probability of drop-out was performed. Possible explanatory variables considered were `Experiment`, `LogSignalStrength` ($\log S_i$), `Cycles` (28, 29, or 30 PCR cycles), `Locus`, `Dye` and `FragmentSize`.

We performed backwards model selection using the Bayesian Information Criterion (Schwarz, 1978) (BIC) to select the best model. The initial model consisted of all first order effects and second order interactions (for example to allow the effect of signal strength to depend on the number of PCR cycles).

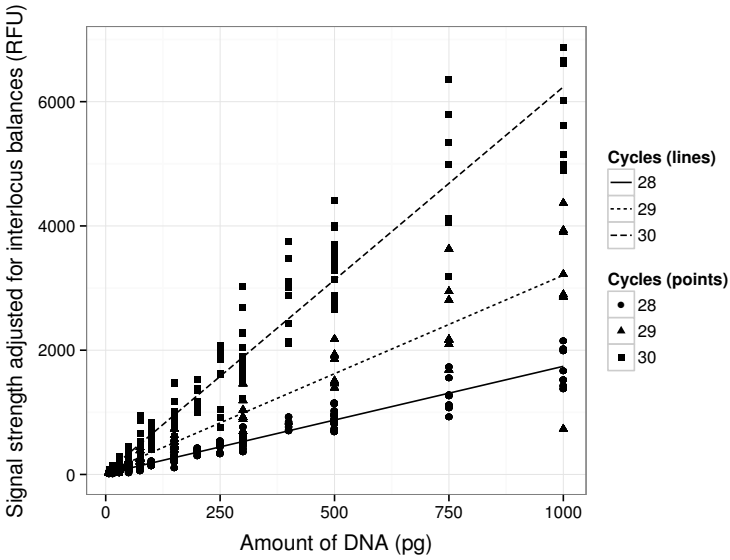


Figure 8. The correlation between the DNA amount and the estimated signal strength (as explained in Section 2.4) given the number of PCR cycles. The lines are linear regression lines for each of the PCR cycles.

3. Results

3.1. *Model for drop-out probability*

As described in Section 2.5, a logistic regression was used to estimate the probability of drop-out. The resulting model was that the drop-out probability is best described by an effect of `LogSignalStrength` ($\log S_i$), `Cycles`, `FragmentSize` and an interaction effect between `LogSignalStrength` and `Cycles` such that the effect of signal strength varies with the number of PCR cycles.

The drop-out probability given signal strength for fragment size 150 bp is shown in Figure 10.

The corresponding signal strength given a drop-out probability for fragment sizes 150 and 300 bp is shown in Figure 11. Table 1 shows the figures.

3.2. *Model validation*

To validate the model, an Hosmer-Lemeshow's test (Hosmer and Lemeshow, 2000) and a bootstrap validation (Breiman, 1996) of the receiver operating characteristic (ROC) were performed.

In total, the dataset contained 6,565 rows (one row per peak). Because of this relatively high number of observations, 50 groups were chosen for the Hosmer-Lemeshow's test. The resulting test statistic was $X^2 = 38.7$, resulting in a non-significant result ($p = 0.83$), meaning that it could not be rejected that the data could be explained by the model.

For a dataset with n samples, the bootstrap procedure was as follows: n

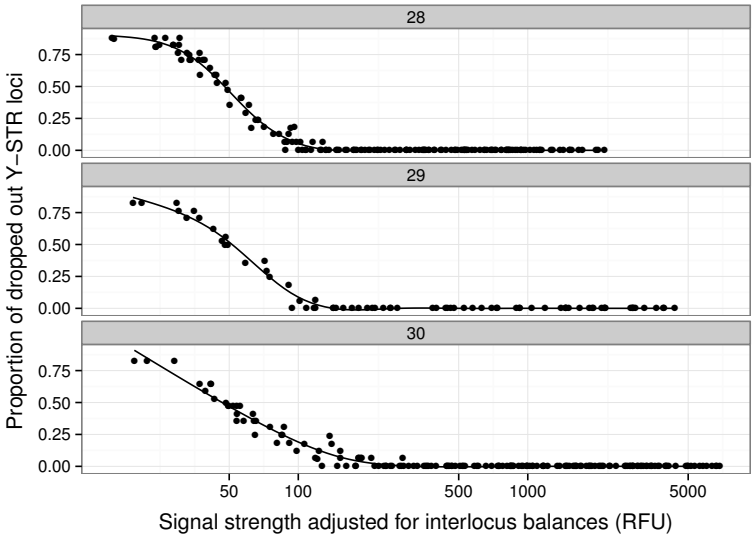


Figure 9. The proportion of dropped out Y-STR loci given signal strength $\theta_j + \log S_i$.

samples were randomly chosen *with* replacement and used to fit the model. The samples from the dataset that were not chosen were then used to validate the model. This was repeated 1,000 times calculating the receiver operating characteristic (ROC). More specifically, the area under the ROC curve (AUC), the sensitivity, and specificity were used as validation statistics. The value of sensitivity and specificity were taken at the cutoff, which was the point, where both were highest with equal weight (meaning that both are treated as equally important, which may not always be the case).

Figure 12 shows the results of the receiver operating characteristic (ROC) analyses of the 1,000 bootstrap realisations. As seen, the results of the ROC analyses did not contradict the proposed model being sufficient to describe the data.

4. Discussion

The result of our investigations indicated that the drop-out probability can be sufficiently described by $\log \hat{S}$ (where $\hat{S}$ is an estimate of the signal strength in a profile), the number of PCR cycles, and fragment size. Note, that the locus balances are incorporated in the calculation of $\log \hat{S}$.

The effects of experiments were not sufficiently strong to be included as a covariate at the model selection, meaning that no significant day-to-day effect was observed. It would be interesting to investigate whether differences in kit-lot number have effect on the parameters under study. Unfortunately, the lot numbers were not recorded.

Going back to the motivating example in Section 1.1, our analysis showed, based on Table 1, that for 30 PCR cycles with a 150 bp allele, the probability of

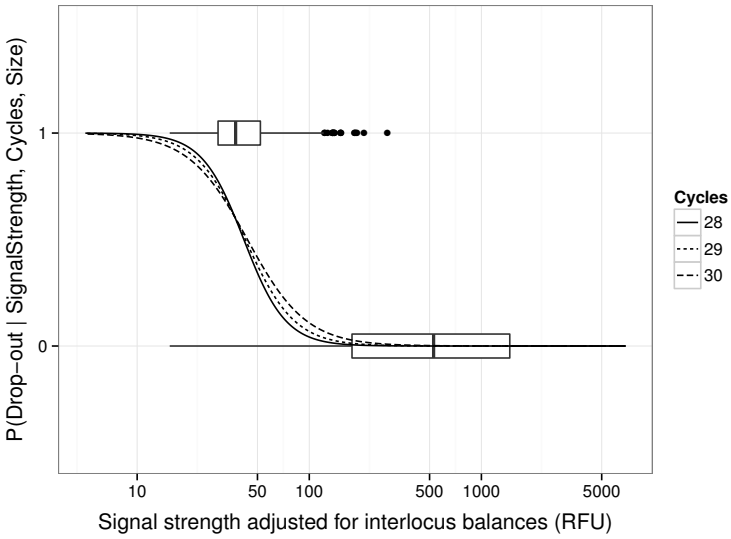


Figure 10. Drop-out probabilities given signal strengths for a fixed fragment size of 150 bp.

drop-out was 1:5,000 corresponding to a rough estimate of the probability of null alleles at a signal strength of S = 1,249 RFU. This means that the probability of a null allele is higher than that of drop-out at 4,000 RFU and that the probability of drop-out is higher than that of a null allele at 75 RFU.

We have developed a model suitable for pristine DNA without degradation. The model can be extended to encompass degraded Y chromosomal DNA similar to the way Tvedebrink *et al.* (2011b) models degraded autosomale DNA.

4.1. *Locus balances*

As already shown in Figure 4, the Yfiler kit is not well balanced. The imbalance seems to be independent of the DNA concentration (not shown). This makes it difficult to make a good model for estimating signal strength.

In Section 2.3, we described a model to estimate the locus balances shown in Figure 5. We will now describe a more advanced model for estimating the signal strength. The idea is that loci with smaller variance contribute with more information to the estimation of the signal strength.

Going back to Figure 4, not all loci have the same variance meaning that they each contribute with a different amount of information. Let ϕ_j^2 be the variance of the j^{th} locus' proportion of the sum of peaks heights (resembles the width of the boxes in Figure 4). As in Equation (1), the full profiles are used to estimate the θ_j's and ϕ_j^2's by using the model

$$\log x_{ij} \sim N\left(\theta_j + \log S_i, \phi_j^2\right). \tag{6}$$

The estimated ϕ_j^2's are depicted in Figure 13. The estimated θ_j's and ϕ_j^2's are

		150 bp			300 bp		
		PCR cycles			PCR cycles		
P(Drop-out)		28	29	30	28	29	30
0.001%	1:100,000	1,050	1,843	4,060	1,296	2,357	5,457
0.002%	1:50,000	865	1,469	3,091	1,067	1,878	4,154
0.01%	1:10,000	551	867	1,640	680	1,109	2,205
0.02%	1:5,000	453	691	1,249	560	884	1,678
0.1%	1:1,000	289	408	663	356	522	891
50%	1:2	42	43	44	51	54	59

Table 1. The signal strength to obtain a given drop-out probability at fragment sizes of 150 bp and 300 bp using a given number of PCR cycles. See Figure 11 for a plot of this table.

then assumed known when used in the model for estimating signal strength, such that

$$\log x_{ij} \sim N_{\log t}\left(\theta_j + \log S_i, \phi_j^2 \sigma_i^2\right), \tag{7}$$

where x_{ij} is the peak height at the j^{th} locus for the i^{th} sample, θ_j is the locus balance for the j^{th} locus and S_i is the signal strength for the i^{th} sample. As seen, Equation (7) is an extension of Equation (2). The likelihood, which for Equation (2) was Equation (4), to be optimised is then

$$L\left(\log S, \sigma^2; \{x_j\}_{j\in J}\right) = \prod_{j\in J^C} \Phi\left(\frac{\log t - (\theta_j + \log S)}{\phi_j \sigma}\right) \times \prod_{j\in J} (\phi_j \sigma)^{-1} \phi\left(\frac{\log x_j - (\theta_j + \log S)}{\phi_j \sigma}\right).$$

The results for the two different ways of estimating signal strength are shown in Figure 14. As seen, the results obtained using the advanced model are quite similar to the results obtained using the simpler model. This does not mean that the variance of the interlocus balances is not important, merely that it is probably difficult to model.

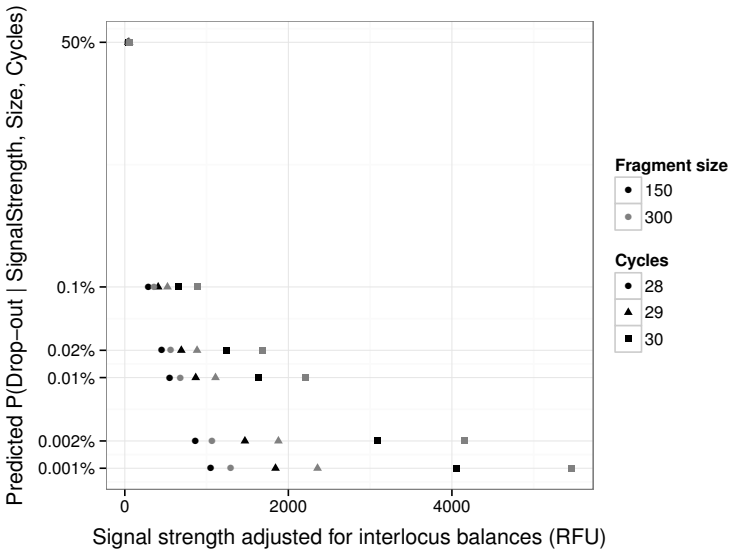


Figure 11. Plot of signal strength given a drop-out probability for fixed fragment sizes 150 and 300 bp. See Table 1 for a table of values used to construct this plot.

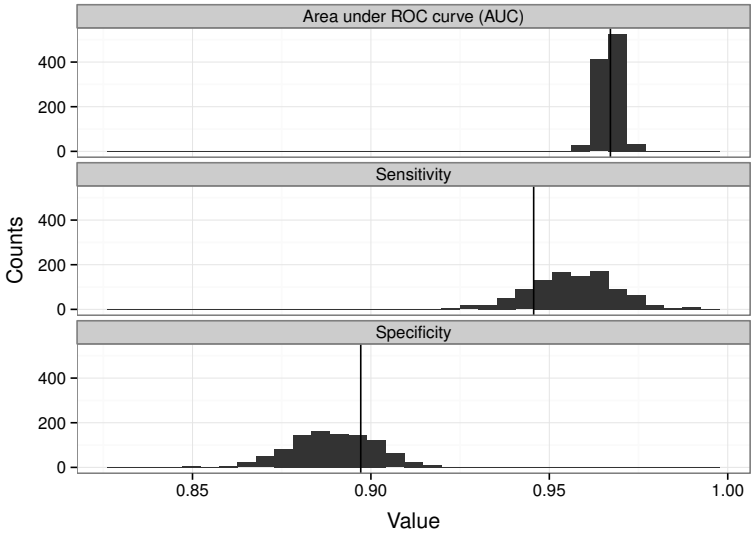


Figure 12. Realisations of the area under curve (AUC), sensitivity and specificity from the ROC analyses of 1,000 bootstrap samples. The vertical lines are the values obtained when both fitting and validating the model using the entire dataset.

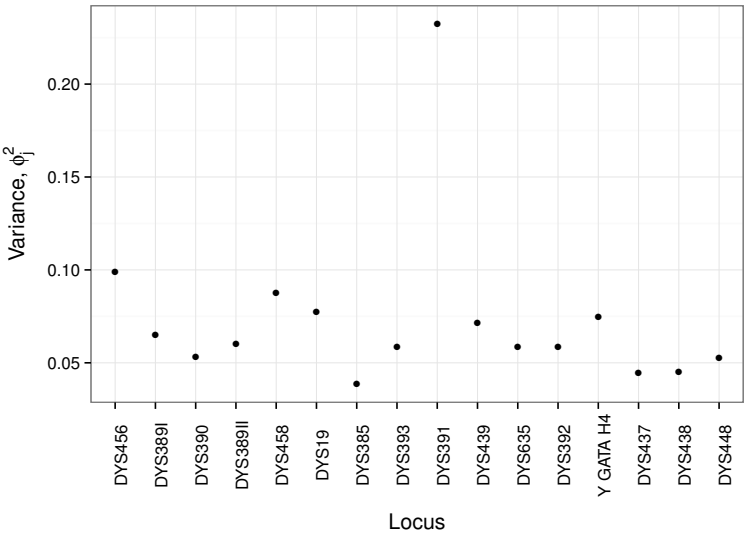


Figure 13. Estimation of the variance, ϕ_j^2, using the model Equation (6). The values can be compared to the width of the boxes in Figure 4. Loci with a large box width in Figure 4 also have a large variance, ϕ_j. One example of this is the DYS391 locus.

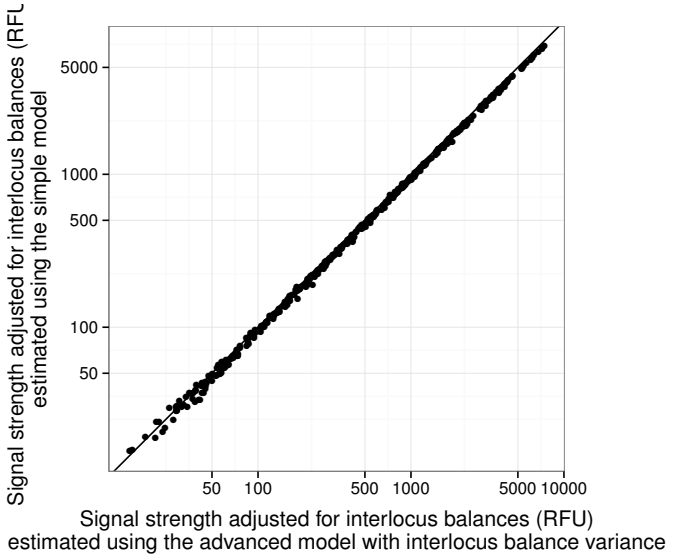


Figure 14. Comparison of the signal strength estimation using the advanced model Equation (7) and the simple model Equation (2). Each point represents the estimated signal strength of a sample using both the advanced and simple model.

5. Bibliography

Agresti, A. (2002) *Categorical Data Analysis*. Wiley, 2. edn. 28

Andersen, M. M., Caliebe, A., Jochens, A., Willuweit, S. and Krawczak, M. (2013) Estimating trace-suspect match probabilities for singleton Y-STR haplotypes using coalescent theory. *Forensic Science International: Genetics*, **7**, 264–271. 20

Breiman, L. (1996) Bagging predictors. *Machine Learning*, **24**, 123–140. 29

Brenner, C. H. (2010) Fundamental problem of forensic mathematics – The evidential value of a rare haplotype. *Forensic Science International: Genetics*, **4**, 281–291. 20

Buckleton, J., Krawczak, M. and Weir, B. (2011) The interpretation of lineage markers in forensic DNA testing. *Forensic Science International: Genetics*, **5**, 78–83. 20

Budowle, B., Aranda, X. *et al.* (2008) Null allele sequence structure at the DYS448 locus and implications for profile interpretation. *International Journal of Legal Medicine*, **122**, 421–427. 20

Butler, J. M. (2005) *Forensic DNA Typing: Biology, Technology, and Genetics of STR Markers*. Academic Press, 2nd edn. 20

Gill, P., Brenner, C., Buckleton, J., Carracedo, A., Krawczak, M., Mayr, W., Morling, N., Prinz, M., Schneider, P. and Weir, B. (2006) DNA commission of the International Society of Forensic Genetics: Recommendations on the interpretation of mixtures. *Forensic Science International*, **160**, 90–101. 20

Gill, P., Brenner, C. *et al.* (2001) DNA Commission of the International Society of Forensic Genetics: Recommendations on forensic analysis using Y-chromosome STRs. *Forensic Science International*, **124**, 5–10. 20

Gusmao, L., Butler, J. *et al.* (2006) DNA Commission of the International Society of Forensic Genetics. DNA Commission of the International Society of Forensic Genetics (ISFG): an update of the recommendations on the use of Y-STRs in forensic analysis. *Forensic Science International*, **157**, 187–197. 20

Hosmer, D. W. and Lemeshow, S. (2000) *Applied Logistic Regression*. Wiley. 28, 29

Krawczak, M. (2001) Forensic evaluation of Y-STR haplotype matches: a comment. *Forensic Science International*, **118**, 114–115. 20

Morling, N., Allen, R., Carracedo, A., Geada, H., Guidet, F., Hallenberg, C., Martin, W., Mayr, W., Olaisen, B., Pascali, V. and Schneider, P. (2002) Paternity Testing Commision of the International Society of Forensic Genetics: recommendations on genetic investigations in paternity cases. *Forensic Science International*, **129**, 148–157. 20

Persson, T. and Rootzen, H. (1977) Simple and Highly Efficient Estimators for a Type I Censored Normal Sample. *Biometrika*, **64**, 123–128. 27

Prinz, M., Boll, K., Baum, H. and Shaler, B. (1997) Multiplexing of Y chromosome specific STRs and performance for mixed samples. *Forensic Science International*, **85**, 209–218. 20

R Development Core Team (2013) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org`. ISBN 3-900051-07-0. 23, 24, 27

Roewer, L. (2009) Y chromosome STR typing in crime casework. *Forensic Sci Med Pathol*, **5**, 77–84. 20

Roewer, L., Kayser, M., de Knijff, P. *et al.* (2000) A new method for the evaluation of matches in non-recombining genomes: application to Y-chromosomal short tandem repeat (STR) haplotypes in European males. *Forensic Science International*, **114**, 31–43. 20

Roewer, L., Krawczak, M., Willuweit, S. *et al.* (2001) Online reference database of European Y-chromosomal short tandem repeat (STR) haplotypes. *Forensic Science International*, **2-3**, 106–113. 20, 21

Schwarz, G. (1978) Estimating the Dimension of a Model. *Annals of Statistics*, **6**, 461–464. 28

Tvedebrink, T., Eriksen, P. S., Asplund, M., Mogensen, H. S. and Morling, N. (2011a) Allelic drop-out probabilities estimated by logistic regression - Further considerations and practical implementation. *Forensic Science International: Genetics*, **6**, 263–267. 20, 25, 28

Tvedebrink, T., Eriksen, P. S., Mogensen, H. S. and Morling, N. (2009) Estimating the probability of allelic drop-out of STR alleles in forensic genetics. *Forensic Science International: Genetics*, **3**, 222–226. 20, 25, 28

Tvedebrink, T., Eriksen, P. S., Mogensen, H. S. and Morling, N. (2011b) Statistical model for degraded DNA samples and adjusted probabilities for allelic drop-out. *Forensic Science International: Genetics*, **6**, 97–101. 31

Willuweit, S. and Roewer, L. (2009) Y chromosome haplotype reference database (YHRD): Update. *Forensic Science International: Genetics*, **1**, 83–87. 20, 21

Part 2

Haplotype distribution modelling

Paper IV

Estimating trace-suspect match probabilities for singleton Y-STR haplotypes using coalescent theory

Author list

Mikkel Meyer Andersen, *Aalborg University, Denmark*
Amke Caliebe, *Christian-Albrechts University, Kiel, Germany*
Arne Jochens, *Christian-Albrechts University, Kiel, Germany*
Sascha Willuweit, *Charité – Universitätsmedizin Berlin, Germany*
Michael Krawczak, *Christian-Albrechts University, Kiel, Germany*

Summary

Estimation of match probabilities for singleton haplotypes of lineage markers, i.e. for haplotypes observed only once in a reference database augmented by a suspect profile, is an important problem in forensic genetics. We compared the performance of four estimators of singleton match probabilities for Y-STRs, namely the count estimate, both with and without Brenner's so-called 'kappa correction', the surveying estimate, and a previously proposed, but rarely used, coalescent-based approach implemented in the BATWING software. Extensive simulation with BATWING of the underlying population history, haplotype evolution and subsequent database sampling revealed that the coalescent-based approach is characterized by lower bias and lower mean squared error than the uncorrected count estimator and the surveying estimator. Moreover, in contrast to the two count estimators, both the surveying and the coalescent-based approach exhibited a good correlation between the estimated and true match probabilities. However, although its overall performance is thus better than that of any other recognized method, the coalescent-based estimator is still computation-intense on the verge of general impracticability. Its application in forensic practice therefore will have to be limited to small reference databases, or to isolated cases of particular interest, until more powerful algorithms for coalescent simulation have become available.

Publication info

This paper was published as:

> Andersen MM, Caliebe A, Jochens A, Willuweit S, Krawczak M (2013). *Estimating trace-suspect match probabilities for singleton Y-STR haplotypes using coalescent theory.* Forensic Science International: Genetics; **7**(2):264-271.

The version here is the journal version with minor typographical corrections and a note about that the modified BATWING program that was made available as the `R` package `rforensicbatwing` (Andersen and Wilson, 2013).

1. Introduction

In forensic genetics, it is often necessary to compare the plausibility of two case-relevant hypotheses on the basis of some genetic data, and the most consistent (and therefore generally recommended) way of doing so is by means of the likelihood ratio (Evett and Weir, 1998). Calculating the likelihood ratio in forensic case work is usually tantamount to quantifying the match probability between two genetic profiles under different assumptions about their degree of relatedness. One particularly important match probability in this context is the probability that a certain individual (e.g. the donor of a trace found at a crime scene) has the same DNA profile as another individual (usually a suspect) drawn randomly from the same population. Methods to estimate this so-called 'trace-suspect' match probability are well established for autosomal STRs (Balding and Nichols, 1994), with most of them assuming statistical independence between the markers included in the profile.

Lineage markers, such as Y-chromosomal short tandem repeats (Y-STRs) or mtDNA polymorphisms, have several advantages over autosomal markers (Gill *et al.*, 1985; Roewer, 2009), for example, when solving cases of sexual assault (Sibille *et al.*, 2002). However, due to the lack of recombination and, therefore, lack of statistical independence, the calculation of match probabilities is more challenging for lineage than for autosomal markers (Buckleton *et al.*, 2011). In particular, when considering Y-STR haplotypes comprising up to 17 loci (Willuweit and Roewer, 2009), the proportion of cases involving singletons, defined as haplotypes observed only once in a reference database augmented by the suspect profile, may become so large that use of traditional count estimates of the corresponding match probabilities becomes unsatisfactory.

To detail the inference problem arising with singleton haplotypes, let us assume that a reference database of size n is given, and that a trace and suspect carry a new haplotype not yet observed in the database. Initially, the count estimator $1/(n+1)$ was used to derive match probabilities in such cases. However, this estimator is rather conservative because it is limited from below by the inverse of the database size. Therefore, a more advanced method referred to as 'haplotype surveying' was proposed by Roewer *et al.* (2000); Krawczak (2001) that tried to exploit the information about evolutionary relatedness inherent in a given database of Y-STR haplotypes. In view of the criticisms raised against it by Andersen (2010); Brenner (2010), the surveying method was later refined by Willuweit *et al.* (2011) and a new version is now implemented, for example, at the YHRD website (Roewer *et al.*, 2001; Willuweit and Roewer, 2009) (see `http://www.yhrd.org`). Recently, Charles Brenner suggested an alternative, comparatively simple method of estimating the match probability for singletons for any kind of markers (Brenner, 2010), the so-called 'κ correction' of the count estimator inspired by Robbins (1968). In short, the κ correction entails estimating a match probability by $(1-\kappa)/(n+1)$, where $\kappa = \alpha/(n+1)$ and α denotes the total number of singletons in the database.

Interestingly, there is yet another estimator of forensic match probabilities that unfortunately never got much attention, most probably due to its computational demands. The approach was first described by Ian Wilson and colleagues in

2003 (Wilson *et al.*, 2003) and involves the refinement of a previously published Markov Chain Monte Carlo method to sample coalescent trees (Kingman, 1982; Wilson and Balding, 1998; Hein *et al.*, 2005). In the present paper, we will briefly recall the original work by Wilson *et al.* (2003) before comparing it to the other three estimators mentioned above. Using both simulated and real data, we will highlight the power and limits of coalescent-based estimation of match probabilities for singleton Y-STR haplotypes.

2. Coalescent-based estimation of match probabilities

The main idea of the coalescent-based approach is as follows (Kingman, 1982; Hein *et al.*, 2005): Adopting a sensible population history and an appropriate mutation model, a large number of coalescent trees is simulated linking the haplotypes in the reference database $H = (h_1, h_2, \ldots, h_n)$ to one another and to the suspect haplotype h_S. Then, the unknown trace donor X is linked randomly to each tree assuming the same population history as in the simulation of the tree. After the tree-specific probabilities have been calculated that the trace donor possesses the same DNA profile as the suspect, the average of these probabilities, taken over all simulated trees, serves as an estimate of the sought-after match probability.

Wilson and Balding (1998) introduced a Bayesian Markov Chain Monte Carlo model to generate random coalescent trees according to their probability of occurrence. This model was expanded in 2003 to include population growth, among other generalizations (Wilson *et al.*, 2003). To our knowledge, the 2003 paper was also the first one to put the calculation of forensic match probabilities into a coalescent theory context: "*In addition to the genealogical tree underlying the* $n+1$ *observed [haplotypes], we introduce a branch connecting the unobserved [haplotype] of [a random individual]* X *with the tree, writing* Z *for the new node thus introduced*". In our terminology, individual Z is the most recent common ancestor of trace donor X and the most closely related individual(s) in the database, including the suspect. In the Bayesian approach taken by Wilson and Balding (1998); Wilson *et al.* (2003), the haplotypes are assumed to be known at all internal nodes of the tree, including h_Z. This implies that the match probability for a given tree equals the probability that h_Z mutates to the suspect haplotype h_S during the time span separating Z and X (Figure 1).

The approach proposed by Wilson *et al.* (2003) is implemented in the computer program 'Bayesian Analysis of Trees With Internal Node Generation' (BATWING), which is publicly available at `http://www.mas.ncl.ac.uk/~nijw/`. However, the BATWING program does not explicitly support the calculation of forensic match probabilities but had to be adapted to this task for the present study. The modified BATWING program with the forensic match probability module included can be downloaded from the 'Software' page at `http://people.math.aau.dk/~mikl/?p=software`. Note, that after publication, the modified BATWING program was also made available as the `R` package `rforensicbatwing` (Andersen and Wilson, 2013)

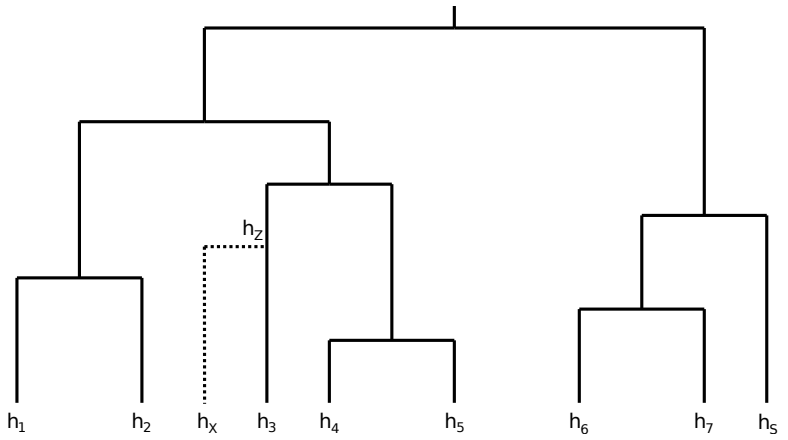


Figure 1. Calculation of forensic match probabilities using coalescent theory (after Wilson *et al.* (2003)). $h_1, h_2, \ldots, h_7$: haplotypes in a reference database H of size $n = 7$; h_S: suspect haplotype; h_X: haplotype of trace donor X; h_Z: haplotype of the most recent common ancestor Z of trace donor X and the most closely related individual(s) in the database, including suspect S. The contribution to the match probability of this particular tree would be the probability that h_Z mutates to h_S during the time span indicated by the dotted line, thereby creating a match between the suspect and trace haplotype.

at `http://cran.r-project.org/package=rforensicbatwing`.

2.1. *Branch-wise contribution to the tree probability*

Calculating match probabilities with BATWING is based upon use of the probabilities that a given haplotype mutates to another given haplotype within a specified period of time. In principle, any realistic mutation model can be employed to quantify these probabilities but, in the case of Y-STRs, it appears reasonable to draw upon a single-step mutation model. Under the single-step mutation model used here, the marker-specific numbers of upward and downward mutations (by one repeat unit) in a given number of generations, M_u and M_d, follow independent Poisson distributions with parameters λ_u and λ_d. For the consequent allelic change, only the net effect of the two opposite mutation processes is important, and this difference, $\Delta = M_u - M_d$, follows a Skellam distribution (Skellam, 1946) with probability function

$$f(\delta; \lambda_u, \lambda_d) = e^{-(\lambda_u + \lambda_d)} \left(\frac{\lambda_u}{\lambda_d}\right)^{\delta/2} I_{|\delta|}(2\sqrt{\lambda_u \lambda_d}).$$

Here, $I_{|\delta|}$ is the modified $|\delta|$th order Bessel function of the first kind. For the sake of simplicity, we will henceforth assume that upward and downward mutations occur at the same rate. In this case, $\lambda = \lambda_u = \lambda_d$ and the Skellam probability function simplifies to

$$f(\delta; \lambda) = e^{-2\lambda} I_{|\delta|}(2\lambda).$$

Now, let N be the effective population size appropriate for a given forensic context, and let $\theta = 2N\mu$ where μ denotes the total mutation rate per generation per marker. Then, the expected number of (upward plus downward) mutations occurring on a tree branch of length t equals $t\theta/2 = tN\mu$. Assuming equal rates for upward and downward mutation, the mutation process can be thought of as creating two independent random variables, each with a Poisson distribution with parameter $(t\theta/2)/2 = t\theta/4$. In summary, the net allelic change $\Delta_t = M_{u,t} - M_{d,t}$ along a tree branch of length t generations thus follows a Skellam distribution with probability function

$$f(\delta; t, \theta) = e^{-t\theta/2} I_{|\delta|}(t\theta/2). \tag{1}$$

2.2. *Estimation of the match probability*

For a given tree, let t denote the time (in generations) between (i) trace donor X and (ii) the most recent common ancestor Z of X and the most closely related individual(s) in the database, including the suspect (Figure 1). As was noted above, the conditional match probability $P(h_X = h_S \mid H, h_S, h_Z, t)$ equals the probability that h_Z mutates into h_S when passed down from Z to X. Since all trees are simulated (approximately) independently according to their conditional probability of occurrence, given reference database H and suspect haplotype h_S, the sought-after match probability $P(h_X = h_S \mid H, h_S)$ can be estimated by

$$\hat{p}_{H,h_S,m} = m^{-1} \sum_{i=1}^{m} P(h_X = h_S \mid H, h_S, h_Z(i), t(i)), \tag{2}$$

where m equals the number of simulated trees, and where $h_Z(i)$ and $t(i)$ refer to the ith tree.

Under the single-step mutation model used here, the conditional probability $P(h_X = h_S \mid H, h_S, h_Z, t)$ can be quantified using the Skellam probability function given in Equation (1). Let $\delta(j) = h_S(j) - h_Z(j)$ be the allelic change required at the jth out of r markers. Then

$$P\left(h_X(j) = h_S(j) \mid H, h_S, h_Z, t\right) = f(\delta(j); t, \theta)$$

and, because of independence between mutations,

$$P(h_X = h_S \mid H, h_S, h_Z, t) = \prod_{j=1}^{r} f(\delta(j); t, \theta).$$

It is worthy of note that coalescent trees are simulated (approximately) independently and according to the same distribution. Therefore, the average of the resulting conditional probabilities $P(h_X = h_S \mid H, h_S, h_Z(i), t(i))$, taken over all m simulations, automatically constitutes a maximum likelihood estimate of the sought-after match probability $P(h_X = h_S \mid H, h_S)$ under the employed coalescent and mutation model.

2.3. *Convergence issues*

The simulation of coalescent trees as described above entails (at least) two different types of convergence of the ensuing match probability estimates:

(i) For a given reference database H and a given suspect haplotype h_S, estimates $\hat{p}_{H,h_S,m}$ from Equation (2) converge to $P(h_X = h_S \mid H, h_S)$ when the number of simulations m increases.

(ii) $P(h_X = h_S \mid H, h_S)$ converges to the true match probability $P(h_X = h_S)$ when the reference database H expands towards the whole population.

This means that, in a given case and with a given reference database, increasing the number of simulations ensures that the coalescent-based estimate of the match probability converges to $P(h_X = h_S \mid H, h_S)$. The latter is an estimate of $P(h_X = h_S)$ and has sampling variance that can only be reduced by increasing the size of the reference database. However, the larger the database, the more simulations would be required for $\hat{p}_{H,h_S,m}$ to approximate $P(h_X = h_S \mid H, h_S)$ sufficiently well, owing to the larger space of coalescent trees to sample from.

3. Methods

The performance of the coalescent-based estimator of singleton match probabilities was compared to that of three other methods, namely (i) the count estimator $1/(n+1)$, where n denotes the database size, (ii) the surveying method in its most recent form (Willuweit *et al.*, 2011), and (iii) Brenner's κ correction of the count estimator (Robbins, 1968; Brenner, 2010).

Each estimator was evaluated on singleton haplotypes from both simulated and real Y-STR data. Simulated data allow a comparison to be made between estimated and true match probabilities by first simulating a big population from which realistically sized databases are then drawn for estimation. As performance measures, we employed the bias and mean squared error (MSE) of each estimator as well as the correlation between the estimated and the truly underlying match probabilities.

Let $\hat{p}_{H_j,h_{S_j}}$ be any estimate of the match probability (coalescent-based, count or surveying) assuming that the jth singleton h_{S_j}, out of v singletons considered, belongs to the suspect. Thus, H_j is the database with the jth singleton excluded. Let $p_{h_{S_j}}$ be the population frequency of h_{S_j} which, for the sake of simplicity, was taken to coincide with the match probability in our study (i.e. the underlying population was assumed to be panmictic). Then the bias of the estimator was estimated by

$$\frac{1}{v}\sum_{j=1}^{v}\left(\hat{p}_{H_j,h_{S_j}} - p_{h_{S_j}}\right).$$

Similarly, the mean squared error was estimated by

$$\frac{1}{v}\sum_{j=1}^{v}\left(\hat{p}_{H_j,h_{S_j}} - p_{h_{S_j}}\right)^2.$$

Finally, we also calculated the Spearman rank correlation coefficient between $\hat{p}_{H_j,h_{S_j}}$ and $p_{h_{S_j}}$. All analyses were carried out with R (R Development Core Team, 2013).

3.1. *Generation and analysis of simulated data*

BATWING by Wilson and Balding (1998); Wilson *et al.* (2003) was not only used for the estimation of match probabilities but also for simulating a large population from which small databases of size $n = 100$ and $n = 200$ were repeatedly sampled for the evaluation of the different estimators. In principle, BATWING supports three different types of population dynamics, namely a constant population size and two exponential growth models (one with constant growth and one with growth after some point in time (Wilson *et al.*, 2003)). Here, we simulated a single source population of 50 million haplotypes that resulted from the constant exponential expansion, over 2,000 generations, of an initial population of 20,000 haplotypes. The two-sided (single-step) mutation rate μ was set equal to 0.003 per generation per marker. The number of markers was set equal to 7 as a compromise between computational feasibility and the possibility to obtain realistic data. As can be inferred from Figure A.1 in Appendix A, the computation time required for coalescent-based match probability estimation for a fixed number of simulations increased dramatically with both the marker number and the database size.

Sample	$n = 100$		$n = 200$	
1	84	(84.0 %)	148	(74.0 %)
2	85	(85.0 %)	135	(67.5 %)
3	82	(82.0 %)	133	(66.5 %)
4	82	(82.0 %)	131	(65.5 %)
5	92	(92.0 %)	152	(76.0 %)

Table 1. Number and percentage (in brackets) of singletons observed in ten databases of different size n, sampled from a large simulated source population. Sample numbers are consistent across Tables 1, 2 and 3.

For each database size (i.e. $n = 100$ or $n = 200$), five databases were drawn randomly from the simulated source population. Next, the forensic match probability was estimated for each singleton haplotype in the database (for the respective proportions of singletons, see Table 1) assuming that the haplotype came from a suspect and was not included in the reference database itself. Estimation was based upon either 500,000 ($n = 100$) or 200,000 ($n = 200$) simulated coalescent trees per singleton. The larger the database, the larger is the space of coalescent trees to sample from. This means that, in principle, more simulations should be performed for larger databases. Due to computational constraints, however, a substantial increase of the simulation number was not feasible in our study. We therefore conducted a partial in-depth analysis for the five databases of size $n = 200$ by randomly selecting 10 singletons from each database and simulating one million trees for each of these.

In the coalescent-based estimation of the match probabilities with BATWING, we used the same distributions of population size, growth rate and mutation rates as employed in the simulation of the source population. This was done in order to verify whether coalescent-based estimation was feasible at all. In practice, such population and mutation parameters may not be known. However, BATWING (Wilson *et al.*, 2003) allows the specification of locus-specific prior distributions that would enable meaningful application of the coalescent-based approach even in cases of uncertainty about the parameters (see subsection "Real data" below).

BATWING's thinning parameters `Nbetsamp` and `treebetN` were both set equal to 15 after minor initial calibration (see the BATWING documentation for further details).

3.2. *Real data*

We analysed the 1,774 German 17-loci haplotypes from release 37 of the YHRD (`http://www.yhrd.org`) (Willuweit and Roewer, 2009). To render the data amenable to both coalescent-based estimation and frequency surveying, some markers and haplotypes had to be excluded. Thus, `DYS385a/b` was ignored because of its inherent genotype ambiguity (Roewer *et al.*, 2000), leaving 15 markers for further analysis. Next, four haplotypes with two alleles reported at `DYS19` and 13 haplotypes with intermediate alleles were excluded, leaving $n = 1{,}757$ haplotypes in the data set. Finally, alleles at `DYS389II` were replaced by `DYS389II` minus `DYS389I` (Butler, 2005). Of the 1,757 haplotypes analysed, 1,469 were singletons (83.6 %).

When restricting the genotype information to the 7-loci so-called 'minimal haplotype' comprising `DYS19`, `DYS389I`, `DYS389II`, `DYS390`, `DYS391`, `DYS392`, and `DYS393`, a total of 392 singletons (22.3 %) were observed in the German data. Ten singletons were drawn randomly from the database and the match probability estimates obtained with the different estimators were compared.

Coalescent-based match probabilities were estimated from 5 million simulations per singleton, after a 50,000 simulations burn-in of the Monte Carlo Markov Chain. All estimations were carried out assuming exponential growth with a `Gamma(1,1)` prior on the growth rate (Wilson *et al.*, 2003), no migration, a `Gamma(3,0.0001)` prior on the effective population size, and fixed mutation rates from `http://www.yhrd.org` as of September 26th, 2012 (`DYS19`: 0.002299, `DYS389I`: 0.002523, `DYS389II`: 0.003644, `DYS390`: 0.002102, `DYS391`: 0.002599, `DYS392`: 0.004123, `DYS393`: 0.001045).

The same thinning parameters as for the simulated data were used (i.e. both `Nbetsamp` and `treebetN` were set equal to 15).

3.3. *Frequency surveying*

Let n_i be the number of times the ith haplotype has been observed in the database including the suspect profile, with $n = \sum_i n_i$ equal to the size of this augmented database and let d_{ij} be the minimum number of mutational steps

separating the ith from the jth haplotype. In its revised form, haplotype surveying (Willuweit *et al.*, 2011) is based upon an exponential regression model

$$\mu_i = \exp(r_1 W_i + r_2),$$
$$\sigma_i = \exp(s_1 W_i + s_2),$$

that links mean μ_i and standard deviation σ_i of the population frequency of the ith haplotype to the weighted inverse molecular distance, $W_i = n^{-1} \sum_{j \neq i} \frac{n_j}{d_{ij}}$, between this haplotype and all other haplotypes in the database. Once the regression parameters r_1, r_2, s_1 and s_2 have been determined, the model serves to define a prior Beta distribution for the frequency of any haplotype h_0 with inverse distance value W_0. The parameters for this prior distribution are

$$\alpha_0 = \frac{\mu_0^2(1-\mu_0)}{\sigma_0^2} - \mu_0,$$
$$\beta_0 = \alpha_0 \left(\frac{1-\mu_0}{\mu_0} \right).$$

Maximum likelihood estimates of the regression parameters were obtained in our study by numerical optimization (Willuweit *et al.*, 2011) using the Nelder-Mead simplex algorithm with up to 1,500 iterations, as implemented in `R` (R Development Core Team, 2013). Several different starting values of (r_1, r_2, s_1, s_2) were tried, and the vector resulting in the highest likelihood was chosen. For the simulated data, starting values were taken from the Cartesian product $\{15, 20\} \times \{-10, -15\} \times \{15, 20\} \times \{-10, -15\}$, resulting in 16 possible vectors to choose from. For the real data, starting values were taken from $\{15, 20, 30.82\} \times \{-10, -15, -13.17\} \times \{15, 20, 28.95\} \times \{-10, -15, -11.71\}$. The additional elements for the real data are the respective binning estimates for the Western-European population adopted from Table 3 of Willuweit *et al.* (2011).

For comparison to the other estimators, we used the mean of the posterior Beta$(\alpha_i + n_i - 1, \beta_i + n - n_i)$, given by

$$\frac{\alpha_i + n_i - 1}{\alpha_i - 1 + \beta_i + n},$$

as the haplotype surveying estimate of the sought-after match probability for h_i. Note that $n_i = 1$ as far as singletons were concerned.

4. Results

4.1. *Comprehensive analysis of all singletons*

Figure 2 illustrates a comparison of the different singleton match probability estimators for database size $n = 100$. Obviously, both the uncorrected count estimator $1/(n+1)$ and the surveying estimator are rather conservative in that almost all estimates were larger than the corresponding true match probability. Brenner's and the coalescent-based estimator, on the other hand, yielded

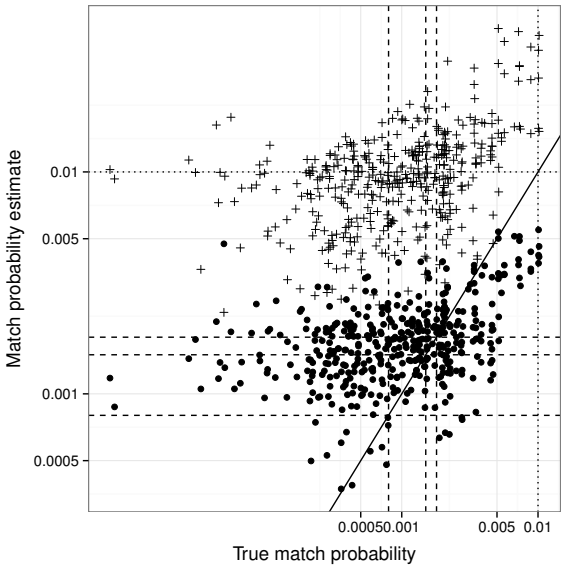


Figure 2. Singleton match probability estimates for five sample databases of size $n = 100$. The uncorrected count estimate (dotted line) was $1/(99+1) = 0.01$ throughout whereas Brenner's estimate varied between 0.0008 and 0.0018, with a mean of 0.0015 (dashed lines). Crosses: surveying estimates; dots: coalescent-based estimates. Each point corresponds to one singleton haplotype. The solid line equates the estimated with the true match probability (i.e. the underlying population frequency).

consistently lower estimates and were found to have small and comparable bias. However, whereas the coalescent-based and surveying estimates were moderately correlated with the true match probabilities, by definition, no such relationship exists for the two count estimators (uncorrected and Brenner's).

Inspection of Figure 2 also reveals that, for singletons with a true match probability smaller than the average of Brenner's estimates, this probability may be difficult to assess by the coalescent-based method in general. On the other hand, singletons with a true match probability above Brenner's average appear to contain sufficient evolutionary information to allow much more precise estimation.

For databases of size $n = 100$ and $n = 200$, Brenner's and the coalescent-based estimator are obviously less biased and have smaller MSE than the surveying estimator (Table 2). In fact, the latter was consistenly found to overestimate the true match probability. Also, for $n = 100$, the coalescent-based estimator had lower MSE than Brenner's. This relationship became reverted for $n = 200$ (Table 2), but there is good reason to believe that this observation essentially reflects insufficient convergence of the coalescent-based estimator because MSE is also a function of the variance of an estimator. As was mentioned above, inspection of the Spearman rank correlation coefficients revealed a moderate correlation with the true match probability for both the surveying and the coalescent-based estimates (Table 2). The correlation between the coalescent-based estimates

	Bias			MSE			Spearman		
	Brenner	Surveying	Coalescent	Brenner	Surveying	Coalescent	Brenner	Surveying	Coalescent
$n = 100$									
Sample 1	$-9.4 \cdot 10^{-5}$	$9.1 \cdot 10^{-3}$	$2.9 \cdot 10^{-4}$	$4.3 \cdot 10^{-6}$	$1.5 \cdot 10^{-4}$	$2.7 \cdot 10^{-6}$	0	0.528	0.446
2	$-3.3 \cdot 10^{-4}$	$8.9 \cdot 10^{-3}$	$2.7 \cdot 10^{-4}$	$4.7 \cdot 10^{-6}$	$8.3 \cdot 10^{-5}$	$2.4 \cdot 10^{-6}$	0	0.566	0.509
3	$4.3 \cdot 10^{-4}$	$9.6 \cdot 10^{-3}$	$4.2 \cdot 10^{-4}$	$2.4 \cdot 10^{-6}$	$9.5 \cdot 10^{-5}$	$2.0 \cdot 10^{-6}$	0	0.413	0.327
4	$5.4 \cdot 10^{-5}$	$8.8 \cdot 10^{-3}$	$-4.0 \cdot 10^{-5}$	$3.4 \cdot 10^{-6}$	$9.8 \cdot 10^{-5}$	$2.3 \cdot 10^{-6}$	0	0.401	0.274
5	$-6.4 \cdot 10^{-4}$	$8.1 \cdot 10^{-3}$	$2.5 \cdot 10^{-4}$	$2.5 \cdot 10^{-6}$	$9.6 \cdot 10^{-5}$	$1.9 \cdot 10^{-6}$	0	0.389	0.266
$n = 200$									
Sample 1	$7.7 \cdot 10^{-5}$	$4.1 \cdot 10^{-3}$	$5.6 \cdot 10^{-5}$	$1.8 \cdot 10^{-6}$	$2.7 \cdot 10^{-5}$	$2.3 \cdot 10^{-6}$	0	0.309	0.154
2	$3.5 \cdot 10^{-4}$	$4.9 \cdot 10^{-3}$	$3.3 \cdot 10^{-4}$	$2.6 \cdot 10^{-6}$	$2.6 \cdot 10^{-5}$	$3.8 \cdot 10^{-6}$	0	0.490	0.267
3	$6.1 \cdot 10^{-4}$	$4.2 \cdot 10^{-3}$	$1.4 \cdot 10^{-4}$	$1.8 \cdot 10^{-6}$	$2.5 \cdot 10^{-5}$	$1.9 \cdot 10^{-6}$	0	0.283	0.343
4	$4.9 \cdot 10^{-4}$	$4.7 \cdot 10^{-3}$	$1.2 \cdot 10^{-4}$	$1.7 \cdot 10^{-6}$	$2.7 \cdot 10^{-5}$	$3.3 \cdot 10^{-6}$	0	0.381	0.184
5	$-8.4 \cdot 10^{-5}$	$4.3 \cdot 10^{-3}$	$9.8 \cdot 10^{-5}$	$2.0 \cdot 10^{-6}$	$2.2 \cdot 10^{-5}$	$2.8 \cdot 10^{-6}$	0	0.389	0.250

Table 2. Comparative analysis of singleton match probability estimators. MSE: mean squared error; Spearman: Spearman rank correlation coefficient between estimated and true match probabilities. Sample database numbers are consistent across Tables 1, 2 and 3.

	Bias			MSE		
	2×10^5	10^6	Brenner	2×10^5	10^6	Brenner
$n = 200$						
Sample 1	$-4.1 \cdot 10^{-4}$	$-1.8 \cdot 10^{-4}$	$-4.7 \cdot 10^{-4}$	$6.0 \cdot 10^{-6}$	$6.0 \cdot 10^{-6}$	$8.9 \cdot 10^{-6}$
2	$-6.7 \cdot 10^{-4}$	$-4.6 \cdot 10^{-4}$	$-2.0 \cdot 10^{-4}$	$3.5 \cdot 10^{-6}$	$2.3 \cdot 10^{-6}$	$4.8 \cdot 10^{-6}$
3	$-6.9 \cdot 10^{-4}$	$-5.6 \cdot 10^{-4}$	$1.9 \cdot 10^{-5}$	$1.3 \cdot 10^{-6}$	$1.3 \cdot 10^{-6}$	$1.7 \cdot 10^{-6}$
4	$1.4 \cdot 10^{-3}$	$4.6 \cdot 10^{-4}$	$6.7 \cdot 10^{-4}$	$6.4 \cdot 10^{-6}$	$1.8 \cdot 10^{-6}$	$1.1 \cdot 10^{-6}$
5	$-3.6 \cdot 10^{-4}$	$-3.2 \cdot 10^{-4}$	$-2.8 \cdot 10^{-4}$	$1.2 \cdot 10^{-6}$	$1.2 \cdot 10^{-6}$	$2.5 \cdot 10^{-6}$

Table 3. In-depth analysis for 10 selected singletons per sample database of the coalescent-based estimator of match probabilities, using different numbers of simulations (2×10^5 and 10^6 per singleton). Sample database numbers are consistent across Tables 1, 2 and 3.

and the true match probabilities was also found to increase with the number of simulations performed (Figure 3). The same was true for the bias and MSE, both of which converged when the number of simulations increased (Figures A.2 and A.3 in Appendix A).

4.2. *In-depth analysis of coalescent-based estimates for selected singletons*

Table 3 summarizes an in-depth analysis of the coalescent-based match probability estimates obtained for 10 randomly selected singletons per sample database of size $n = 200$, using a much larger number of simulations than before. In general, a substantial increase in simulation number from 200,000 to one million reduced both bias and MSE. We also generated two individual trace plots of one million simulations and included these into Appendix A. For one singleton (Figure A.4) convergence of the match probability estimate was lacking while, for the other singleton (Figure A.5), the match probability estimate converged quite rapidly.

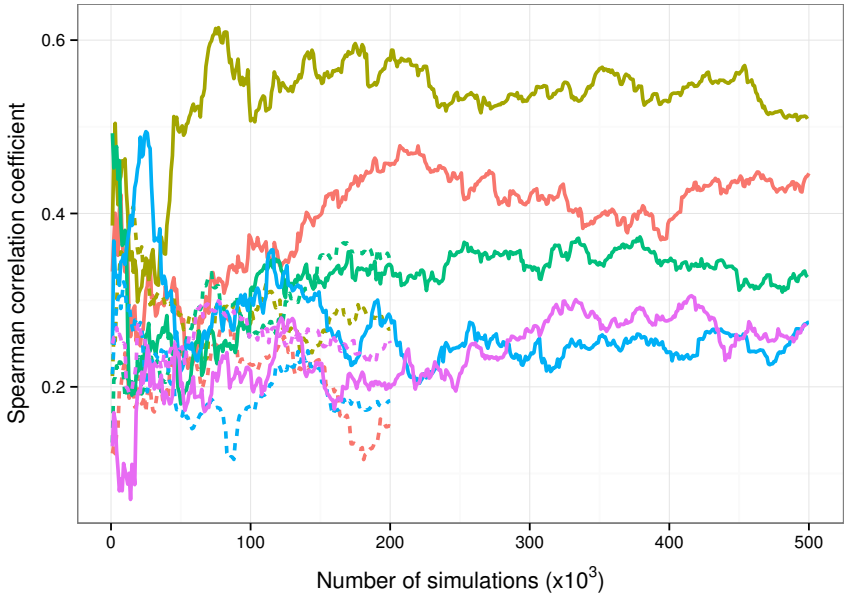


Figure 3. Trace plots of the Spearman rank correlation coefficient between true match probabilities and coalescent-based estimates, after a given number of simulations. Each line corresponds to one of five databases per database size n, sampled at random from a large simulated source population. Solid lines: $n = 100$; dashed lines: $n = 200$.

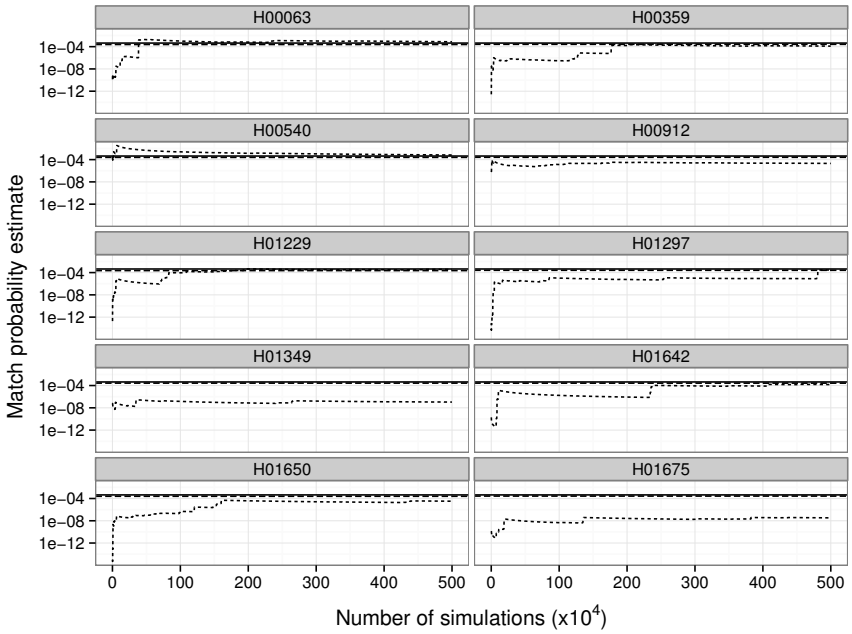


Figure 4. Trace plots of selected match probability estimates from the real German 7-loci Y-STR data. Match probabilities were estimated for 10 singletons using the coalescent-based approach with 5 million simulations (solid lines), Brenner's κ correction (dotted lines) and the surveying estimator (dash-dotted lines). Note: Brenner's estimate equaled 4.0×10^{-4} while the surveying estimate ranged from 2.3×10^{-4} to 2.7×10^{-4} for the 10 haplotypes. Since the vertical axis has a logarithmic scale, the less than two-fold difference between the two types of estimates implied that they were depicted in close proximity. Trace plots of a subsample study of H01675 can be found in Figure A.8 in Appendix A.

4.3. *Real data*

Trace plots for the 7-loci match probabilities of ten singletons randomly chosen from the real German Y-STR data are given in Figure 4. In some instances, but not all, the coalescent-based estimates seem to have converged to a value near Brenner's and the surveying estimates. A singleton that does not seem to have converged at all is H01675. Inspired by Felsenstein (2006), we drew 10 random subsamples of 50 haplotypes each from the original database to see if the subsample estimates for H01675 approximated the other estimates. The trace plots can be found in Figure A.8 in Appendix A. Since the true match probabilities were unknown for the real data, a comparison of the different estimators in terms of their accuracy was not possible. However, the mean subsample estimates for H01675 were in the range of 10^{-2} to 10^{-3}, indicating that the original coalescent-based estimate had indeed not converged, despite the large number of simulations performed.

5. Discussion

5.1. *General appraisal of coalescent-based match probability estimation*

Our simulation study revealed that, overall, the coalescent-based estimator of trace-suspect match probabilities performs better for Y-STR singleton haplotypes than any other previously proposed estimator, at least under the conditions of our simulation study. In terms of both its bias and mean squared error (MSE), the coalescent-based approach was found to be clearly superior to the surveying method by Roewer *et al.* (2000); Krawczak (2001); Willuweit *et al.* (2011). Moreover, it also outperformed the κ correction by Brenner (2010) regarding the correlation between estimated and true match probability which, by definition, equals zero for Brenner's estimator. The said correlation also indirectly corroborated the claim, made in connection with the first introduction of the surveying method (Roewer *et al.*, 2000), that the allelic spectrum of a given database contains valuable information about the evolutionary relatedness of its constituent haplotypes, and therefore about match probabilities.

This view is further supported by the observation that, for all the singletons analysed in our simulation study combined (Table 1), the correlation between the true match probabilities and their coalescent-based estimates increases with the key parameter of the surveying method (Willuweit *et al.*, 2011), namely the weighted inverse molecular distance W between a singleton and the rest of the corresponding reference database (Table 4).

W_i range	No. singletons	Spearman
(0.05, 0.10]	138	0.077
(0.10, 0.15]	451	0.082
(0.15, 0.20]	331	0.130
(0.20, 0.25]	154	0.155
(0.25, 0.40]	50	0.442

Table 4. For each range of W_i values, the Spearman rank correlation coefficient between the coalescent-based estimates and the true match probabilities is given together with the number of singletons in each range.

The major downside of the coalescent-based approach consists in its enormous computational demands. These render any wide-spread practical application of the method difficult, at least until more powerful algorithms to sample coalescent trees have been developed and implemented in suitable software packages. Moreover, because of the large number of singletons assessed in our study, the number of simulations performed for each individual estimate had to be comparatively low. Therefore, the resulting biases and MSEs still have to be interpreted with some caution. This notwithstanding, if applied to derive only one or a few match probability estimates, and with a greater number of simulations thus possible, our in-depth analysis of selected singletons suggests that the accuracy of the coalescent-based method will surpass that of the other approaches tested.

As has been mentioned in Section 3, the Bayesian framework of BATWING

(Wilson *et al.*, 2003) allows the specification of prior distributions for the coalescent parameters, including the effective population size, (locus-specific) mutation rates and population growth. This way, any uncertainty about the respective quantities (as would arise in practical casework) can be incorporated into the population model and the posterior distributions derived. Here, we used fixed mutation rates and standard prior distributions for the other parameter values because our main interest was to determine if and how the coalescent-based method would work in principle. Along the same vein, we employed a simplified mutation model in our study for which the upward and downward mutation rates were assumed to be equal. In practice, if the coalescent-based approach was to be used to estimate real match probabilities, this assumption can be abandoned in favor of allele- and direction-specific mutation rates for the Y-STRs of interest, although such modifications may require substantial alteration of the software used.

To assess the robustness of the coalescent-based estimate, we also varied the mutation rate and the prior distribution of the effective population size. The resulting trace-plots can be found in Figure A.6 and Figure A.7 in Appendix A. With all the different values and priors tested, the coalescent-based estimator turned out to be quite robust.

5.2. *Match probabilities for non-singletons*

In our study, we focused upon singleton haplotypes, i.e. haplotypes for which the estimation of match probabilities appears to be most problematic because the commonly used count estimator $1/(n+1)$ is rather conservative. Moreover, singleton proportions are bound to increase with the number of markers included in a genetic profile, and particularly so when rapidly mutating Y-STRs (Ballantyne *et al.*, 2010) are involved. However, one important advantage of the coalescent-based (and the surveying) estimator over Brenner's κ correction of the count estimate is that singletons are not treated differently from other, more frequent haplotypes. Therefore, the coalescent-based method can be expected to work as reliably for non-singletons as for singletons, although this supposition still needs to be confirmed systematically.

5.3. *Computational recommendations*

The coalescent-based method is still on the verge of being too slow for practical application, at least with the software used here. This is because the computation time required grows exponentially with both the database size and the number of loci involved (Figure A.1). In addition, the more markers are included in a genetic profile, and the larger the database used to quantify the evidential weight of a match, the more simulations are required to guarantee proper convergence of the coalescent-based estimate of the match probability. Therefore, the practical application of the coalescent-based approach would currently be limited to rather small databases and to small numbers of markers.

The above notwithstanding, some recommendations can still be made to

facilitate efficient and sensible use of the existing simulation software. First, when using Metropolis-Hastings sampling (Metropolis *et al.*, 1953; Hastings, 1970) as done in BATWING (Wilson *et al.*, 2003), it is important to carefully choose the acceptance rates so as to ensure that the algorithm visits a sufficiently large proportion of the parameter space. There are guidelines regarding the best choice of proposal functions and acceptance rates (Gelman *et al.*, 1996) and these should be adopted if and when meaningful. Second, thinning parameters such as `Nbetsamp` and `treebetN` should be calibrated to individual cases, for example, by consulting autocorrelation plots and statistics, so that the simulations are made approximately independent. Third, the rate and quality of the convergence of individual estimates should be assessed by trace plots similar to those of Figure 4. Finally, like with other Markov Chain Monte Carlo methods, a burn-in is recommended for the use of BATWING.

6. Acknowledgements

We wish to thank Ian J. Wilson, Newcastle upon Tyne, for providing non-released parts of the forensic match probability extension of the BATWING program and for helping us with implementing them in BATWING. We also wish to thank Charles Brenner, Oakland, for additional comments on (Brenner, 2010) and for a fruitful discussion of the topic in general. Two anonymous reviewers are gratefully acknowledged for helping us to improve our paper.

7. Bibliography

Andersen, M. M. (2010) *Y-STR: Haplotype Frequency Estimation and Evidence Calculation*. Master's thesis, Aalborg University, Denmark. 40

Andersen, M. M. and Wilson, I. J. (2013) *rforensicbatwing: BATWING for calculating forensic trace-suspect match probabilities*. R package version 1.1. 39, 41

Balding, D. J. and Nichols, R. A. (1994) DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands. *Forensic Sci. Int.*, **64**, 125–140. 40

Ballantyne, K. N. *et al.* (2010) Mutability of Y-Chromosomal Microsatellites: Rates, Characteristics, Molecular Bases, and Forensic Implications. *The American Journal of Human Genetics*, **87**, 341–353. 53

Brenner, C. H. (2010) Fundamental problem of forensic mathematics – The evidential value of a rare haplotype. *Forensic Science International: Genetics*, **4**, 281–291. 40, 44, 52, 54

Buckleton, J., Krawczak, M. and Weir, B. (2011) The interpretation of lineage markers in forensic DNA testing. *Forensic Science International: Genetics*, **5**, 78–83. 40

Butler, J. M. (2005) *Forensic DNA Typing: Biology, Technology, and Genetics of STR Markers*. Academic Press, 2nd edn. 46

Evett, I. W. and Weir, B. S. (1998) *Interpreting DNA Evidence*. Sinauer Associates. 40

Felsenstein, J. (2006) Accuracy of Coalescent Likelihood Estimates: Do We Need More Sites, More Sequences, or More Loci? *Mol. Biol. Evol.*, **23**, 691–700. 51

Gelman, A., Robers, G. O. and Gilks, W. R. (1996) Efficient Metropolis Jumping Rules. *Bayesian Statistics*, **5**, 599–607. 54

Gill, P., Jeffreys, A. J. and Werrett, D. J. (1985) Forensic application of DNA fingerprints. *Nature*, **318**, 577–579. 40

Hastings, W. K. (1970) Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika*, **57**, 97–109. 54

Hein, J., Schierup, M. H. and Wiuf, C. (2005) *Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory*. Oxford University Press. 41

Kingman, J. F. C. (1982) The Coalescent. *Stochastic Processes and their Applications*, **13**, 235–248. 41

Krawczak, M. (2001) Forensic evaluation of Y-STR haplotype matches: a comment. *Forensic Science International*, **118**, 114–115. 40, 52

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. and Teller, E. (1953) Equations of State Calculations by Fast Computing Machines. *Genetics*, **21**, 1087–1092. 54

R Development Core Team (2013) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org`. ISBN 3-900051-07-0. 45, 47

Robbins, H. E. (1968) Estimating the Total Probability of the Unobserved Outcomes of an Experiment. *The Annals of Mathematical Statistics*, **39**, 256–257. 40, 44

Roewer, L. (2009) Y chromosome STR typing in crime casework. *Forensic Sci Med Pathol*, **5**, 77–84. 40

Roewer, L., Kayser, M., de Knijff, P. *et al.* (2000) A new method for the evaluation of matches in non-recombining genomes: application to Y-chromosomal short tandem repeat (STR) haplotypes in European males. *Forensic Science International*, **114**, 31–43. 40, 46, 52

Roewer, L., Krawczak, M., Willuweit, S. *et al.* (2001) Online reference database of European Y-chromosomal short tandem repeat (STR) haplotypes. *Forensic Science International*, **2-3**, 106–113. 40

Sibille, I., Duverneuil, C. *et al.* (2002) Y-STR DNA amplification as biological evidence in sexually assaulted female victims with no cytological detection of spermatozoa. *Forensic Sci. Int.*, **125**, 212–216. 40

Skellam, J. G. (1946) The frequency distribution of the difference between

two Poisson variates belonging to different populations. *Journal of Royal Statistical Society: Series A*, **109**, 296. 42

Willuweit, S., Caliebe, A., Andersen, M. M. and Roewer, L. (2011) Y-STR Frequency Surveying Method: A critical reappraisal. *Forensic Science International: Genetics*, **5**, 84–90. 40, 44, 47, 52

Willuweit, S. and Roewer, L. (2009) Y chromosome haplotype reference database (YHRD): Update. *Forensic Science International: Genetics*, **1**, 83–87. 40, 46

Wilson, I. J. and Balding, D. J. (1998) Genealogical Inference From Microsatellite Data. *Genetics*, **150**, 499–510. 41, 45

Wilson, I. J., Weale, M. E. and Balding, D. J. (2003) Inferences from DNA Data: Population Histories, Evolutionary Processes and Forensic Match Probabilities. *Journal of Royal Statistical Society Series A*, **166**, 155–201. 41, 42, 45, 46, 53, 54

Appendix A. Supplementary figures

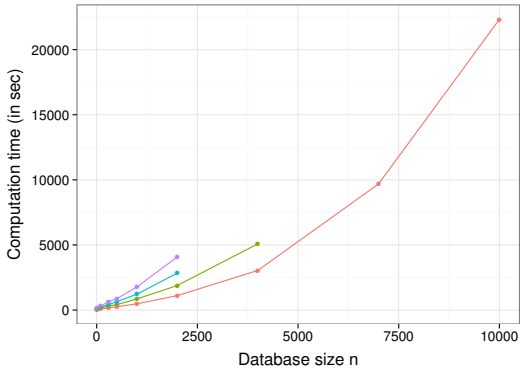


Figure A.1. Computational demand of coalescent-based estimation of singleton match probabilities as a function of database size n and number of loci included in a genetic profile. Calculations were run on an Intel Xeon CPU E5420 at 2.50GHz. Computation times are averages per singleton haplotype. Parameters used were: 10,000 simulations per coalescent tree, a starting population size of 20,000, no population growth, no migration, and a mutation rate of 0.003 per locus per generation. Red dots: 5 loci; green dots: 10 loci; blue dots: 15 loci; black dots: 20 loci.

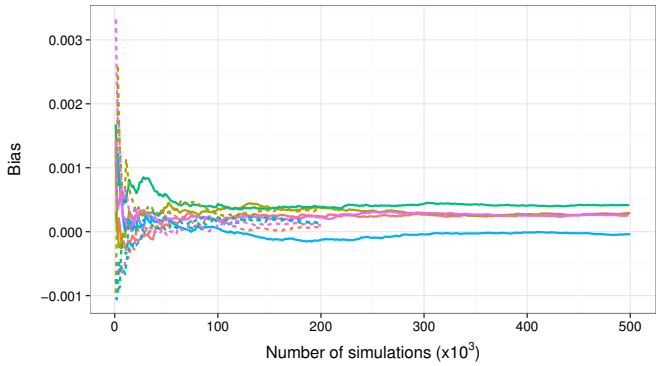


Figure A.2. Bias (see Equation 3 of the main text) of the coalescent-based estimator of singleton match probabilities. Calculation of the bias was based upon all singletons in each of five databases per database size. Solid lines: database size $n = 100$; dashed lines: database size $n = 200$.

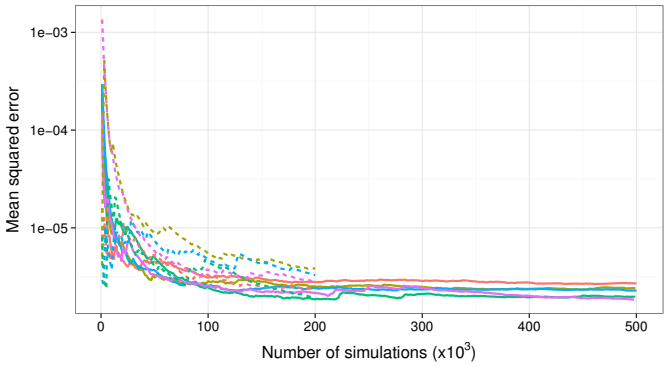


Figure A.3. Mean squared error (see Equation 4 of the main text) of the coalescent-based estimator of singleton match probabilities. Calculation of the MSE was based upon all singletons in each of 5 databases per database size. Solid lines: database size $n = 100$; dashed lines: database size $n = 200$.

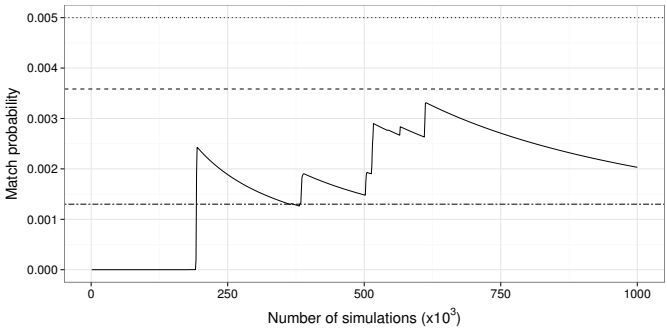


Figure A.4. Trace plot of the coalescent-based match probability estimate (solid line) for a selected singleton from sample database no. 1 of size $n = 200$. Dotted line: uncorrected count estimate; dash-dotted line: Brenner's κ-corrected count estimate, dashed line: true match probability (i.e. the underlying population frequency).

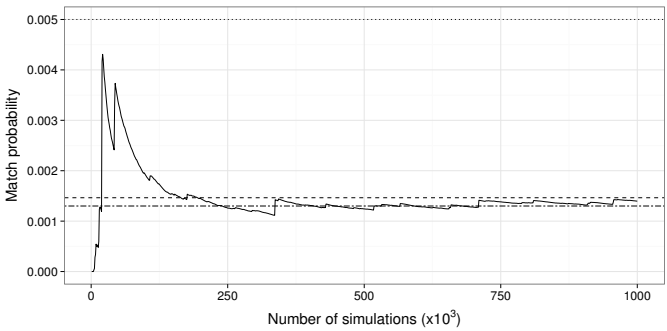


Figure A.5. Trace plot of the coalescent-based match probability estimate (solid line) for a selected singleton from sample database no. 1 of size $n = 200$. Dotted line: uncorrected count estimate; dash-dotted line: Brenner's κ-corrected count estimate; dashed line: true match probability (i.e. the underlying population frequency).

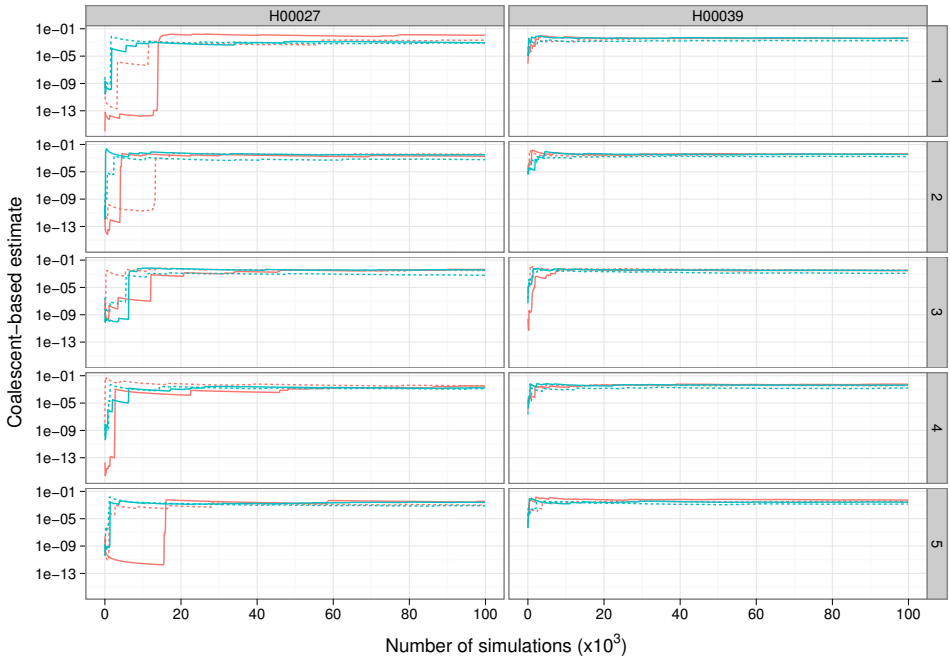


Figure A.6. Trace plots from a robustness study. For each of two randomly selected singletons (columns), five subsamples (rows) of size 20 were drawn from the original database of size 100 (database 1 in Tables 1 and 2 of the main text). Match probabilities were then estimated from each of these subsamples alone, using the coalescent approach. Mutation rates were fixed at either 0.001 (red lines) or 0.003 (blue lines). The effective population size was assumed to be normally distributed with a mean of either 10,000 (solid lines) or 20,000 (dashed lines) and a standard deviation of 3,000. See Figure A.7 for a magnified traceplot where the first 20,000 simulations were discarded as burn-in.

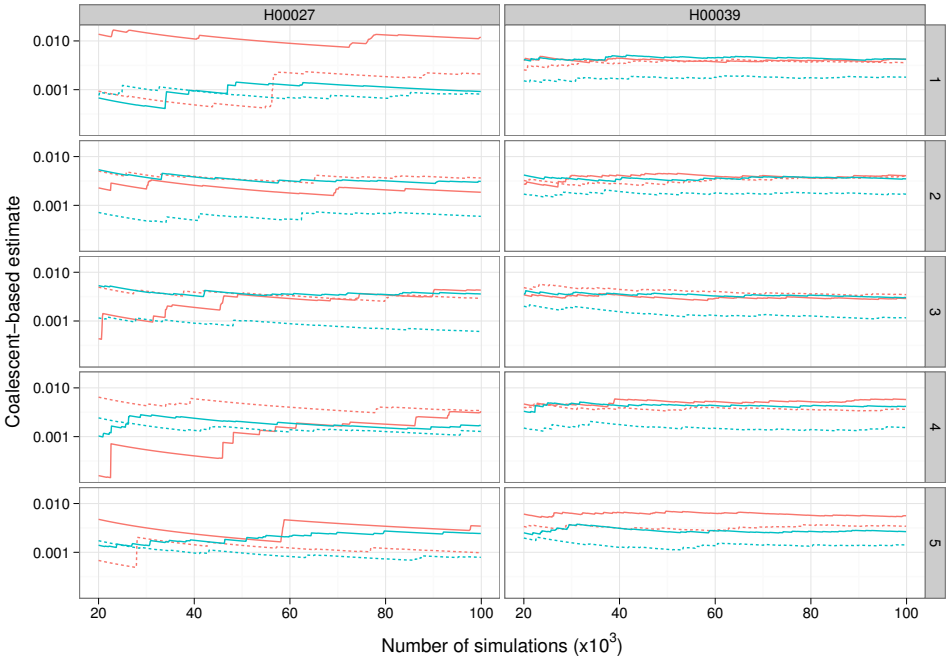


Figure A.7. Same as Figure A.6 but with the first 20,000 simulations discarded as burn-in.

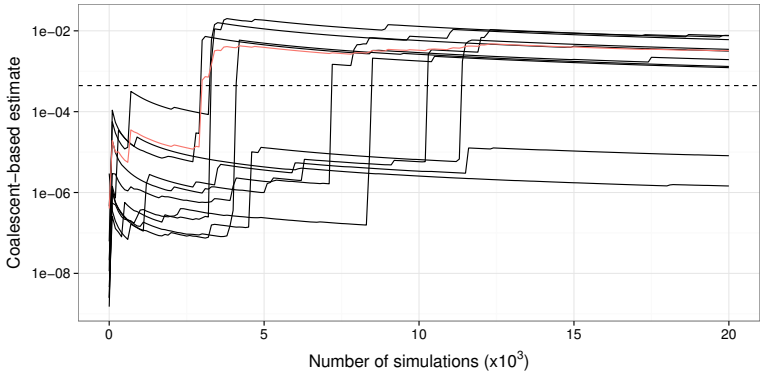


Figure A.8. Subsampling study on singleton haplotype H01675 from the German 7-loci database (1,757 haplotypes). Ten subsamples of 50 haplotypes each were randomly drawn from the database. With each of these subsamples, a coalescent-based estimate was calculated using 20,000 simulations. All estimations were carried out assuming exponential population growth with a `Gamma(1,1)` prior on the growth rate, no migration, a `Gamma(3,0.0001)` prior on the effective population size, and the following mutation rates from `http://www.yhrd.org` as of September 26th, 2012: `DYS19`, 0.002299; `DYS389I`, 0.002523; `DYS389II`, 0.003644; `DYS390`, 0.002102; `DYS391`, 0.002599; `DYS392`, 0.004123; `DYS393`, 0.001045. The black solid lines depict the individual coalescent-based estimation processes. The red solid line depicts the mean of individual runs. The dashed black line corresponds to Brenner's estimate.

Paper V

Efficient forward simulation of Fisher-Wright populations with stochastic population size and neutral single step mutations

Author list Mikkel Meyer Andersen, *Aalborg University, Denmark*
Poul Svante Eriksen, *Aalborg University, Denmark*

Summary In both population genetics and forensic genetics it is important to know how haplotypes are distributed in a population. Simulation of population dynamics helps facilitating research on the distribution of haplotypes. In forensic genetics, the haplotypes can for example consist of lineage markers such as short tandem repeat loci on the Y chromosome (Y-STR). A dominating model for describing population dynamics is the simple, yet powerful, Fisher-Wright model. We describe an efficient algorithm for exact forward simulation of exact Fisher-Wright populations (and not approximative such as the coalescent model). The efficiency comes from convenient data structures by changing the traditional view from individuals to haplotypes. The algorithm is implemented in the open source `R` package `fwsim` and is able to simulate very large populations. We focus on a haploid model and assume stochastic population size with flexible growth specification, no selection, a neutral single step mutation process, and self-reproducing individuals. These assumptions make the algorithm ideal for studying lineage markers such as Y-STR.

Publication info This paper has been made publicly available as:

Andersen MM and Eriksen PS (2012). *Efficient Forward Simulation of Fisher-Wright Populations with Stochastic Population Size and Neutral Single Step Mutations in Haplotypes.* arXiv: 1210.1773.

1. Introduction

Simulation of population dynamics is an important tool when studying genetic traits. In both population genetics and forensic genetics it is important to know how haplotypes are distributed in a population. In forensic genetics, the haplotypes can for example consist of lineage markers such as short tandem repeat loci on the Y chromosome (Y-STR). Simulation of population dynamics helps facilitating research on the distribution of haplotypes. A dominating model for describing population dynamics is the simple, yet powerful, Fisher-Wright model (or process) (Fisher, 1922, 1930, 1958; Wright, 1931; Ewens, 2004). In population genetics, the model also forms the basis for coalescent theory (Kingman, 1982; Hudson, 2001; Hein *et al.*, 2005).

Because the Fisher-Wright model is widely used in population genetics, efficient simulation algorithms and tools are needed. In this paper we describe the model implemented in the `R` (R Development Core Team, 2013) package `fwsim` (Andersen and Eriksen, 2012), which provides an efficient tool for simulating certain kinds of Fisher-Wright populations. The simulation scheme described in this paper is exact (from the Fisher-Wright model) and not approximative like the simulation scheme from the coalescent model (Kingman, 1982; Hudson, 2001; Hein *et al.*, 2005).

Ewens (2004) is a good reference on different models in population genetics as it explains several models and also gives theoretical results.

First some nomenclature must be introduced. Let a locus (loci in plural) be a specific location on the chromosome. The content of a locus is called an allele, which consists of DNA sequences. Here, we assume that the alleles are short tandem repeats (STRs) (Butler, 2005) with values in $\mathbb{Z}$ (in genes, an allele could also just be either of two states, A or B, say). A haplotype is a ordered collection of alleles at loci that are transmitted together.

We focus on a haploid model, where each individual is a gamete with a haplotype consisting of r loci. Hence, a haplotype can in this context be thought of as a vector in $\mathbb{Z}^r$. It may for example be an Y-STR haplotype. We assume no selection and the individuals are self-reproducing.

First, the traditional Fisher-Wright model without mutations is described in order to introduce the notation and to make it possible to compare it with our model.

Throughout this paper, whenever there is a mutation process, we assume it to be a neutral (in the sense of no selection) single step mutation process with infinitely many possible allelic states. This model was introduced by Ohta and Kimura (1973) and some mathematical properties were recently discussed in (Caliebe *et al.*, 2010).

1.1. *Fisher-Wright model without mutation*

Traditionally, a simple Fisher-Wright model, for example as formulated by Ewens (2004), assumes constant population size and no mutations. A Fisher-Wright model is often characterised by a binomial sampling scheme focusing on individuals (or a multinomial sampling scheme focusing on the entire

population), such that a new generation of children is sampled by letting each child choose its parent (and thus its haplotype) uniformly at random.

Because our interest is aimed at the sampling of populations and not at the genealogy, the focus is now changed from individuals to haplotypes, where identical haplotypes are treated similarly, as we are not interested in the genealogical tree itself, but only in the haplotypes and their counts in the resulting population (and possibly in the intermediate populations, too).

Let N be the constant, known population size and H the set of haplotypes. Denote by $n_i(x)$ the number of haplotypes in the i'th generation of haplotype $x \in H$ and $z_{i+1}(x)$ the number of children from haplotype $x \in H$ in generation $i+1$. Because there are no mutations, we have that $n_{i+1}(x) = z_{i+1}(x)$.

The simple Fisher-Wright model arises by assuming that

$$P(\{n_{i+1}(x)\}_{x\in H} \mid \{n_i(x)\}_{x\in H})$$

is given by

$$\{n_{i+1}(x)\}_{x\in H} \mid \{n_i(x)\}_{x\in H} \sim \text{Multinomial}\left(N, \left\{\frac{n_i(x)}{N}\right\}_{x\in H}\right). \tag{1}$$

A property of the multinomial distribution is that

$$\mathbf{E}[n_{i+1}(x) \mid n_i(x)] = n_i(x)$$

as expected.

We note that the process is a Markov chain with $|H|$ absorbing states, one for each haplotype.

2. Model

As mentioned in Section 1.1, the model is formulated on the basis of haplotypes instead of individuals, because it is much more efficient when we are interested in the resulting population after a number of generations rather than the genealogy.

The notation from Section 1.1 is adopted, such that H_i is the set of haplotypes in the i'th generation (H_i depends on i due to mutations, which will be introduced below), $n_i(x)$ is the number of haplotypes in the i'th generation of haplotype $x \in H_i$, and $z_{i+1}(x)$ the number of children from haplotype $x \in H_i$. Now let $N_i = \sum_{x\in H_i} n_i(x)$ be the population size in the i'th generation (instead of a constant population size N as in the simple Fisher-Wright model in Equation (1)).

Our model is then a specification of how

$$\{z_{i+1}(x)\}_{x\in H_i} \mid \{n_i(x)\}_{x\in H_i}$$

is distributed, that is, how the haplotypes in the next generation are conditionally distributed given the previous generation.

Two important features of our model is, that it assumes stochastic population size – which we believe is a more realistic model – and allows flexible population growth specification. We believe that the Fisher-Wright model that will be

introduced below with stochastic population size also incorporating flexible population growth has not yet been defined like we do in the following. First the modelling of the population size and growth will be described. Afterwards the mutational model will be explained.

2.1. *Population size and growth*

Let N_0 be the known initial population size. Note that in the traditional Fisher-Wright model, this is assumed to be a constant.

Then we assume that

$$N_i \mid N_{i-1} \sim \text{Poisson}(\alpha_i N_{i-1}) \tag{2}$$

for $\alpha_i > 0$ ($\alpha_i > 1$ gives growth and $0 < \alpha_i < 1$ gives decline). For example, if $\alpha_i = \alpha$ for all i, then

$$\mathbf{E}[N_i] = \alpha^i N_0,$$

that is exponential population growth. One could also choose

$$\alpha_i = \begin{cases} \beta, & \text{for } i \leqslant t, \\ \alpha, & \text{else,} \end{cases}$$

yielding

$$\mathbf{E}[N_i] = \begin{cases} \beta^i N_0, & \text{for } i \leqslant t, \\ \beta^t \alpha^{i-t} N_0, & \text{else,} \end{cases}$$

which for example can be used to get exponential growth up to generation t and afterwards an expected constant population size by setting $\alpha = 1$.

A possibly more realistic example is logistic population growth, which can be obtained by specifying a maximum population size N_{max}, $\alpha \geqslant 1$, and then setting

$$\alpha_i = \alpha - \frac{(\alpha - 1) N_{i-1}}{N_{max}}$$

as the growth rates. A closed form expression for $\mathbf{E}[N_i]$ in this case seems difficult to obtain.

One could alternatively also create a (possibly decreasing) rate $\alpha_i = f(i)$ for some function f. Hence, the specification of growth is rather flexible.

2.2. *Number of children*

As mentioned previously, the conditional distribution $\{z_{i+1}(x)\}_{x \in H_i} \mid \{x_i(x)\}_{x \in H_i}$ must be specified. We assume that the number of children $z_{i+1}(x_0)$ of a certain haplotype $x_0 \in H_i$ is conditionally independent of the number of children of other haplotypes, given the entire previous generation $\{x_i(x)\}_{x \in H_i}$. Thus, only the marginal distribution $z_{i+1}(x_0) \mid \{x_i(x)\}_{x \in H_i}$ must be specified.

For each haplotype $x_0 \in H_i$ in the i'th generation occuring $n_i(x_0)$ times, we then assume that the number of children $z_{i+1}(x_0)$ is distributed independently of other haplotypes as

$$z_{i+1}(x_0) \mid \{n_i(x)\}_{x \in H_i} \sim \text{Poisson}(\alpha_{i+1} n_i(x_0)). \tag{3}$$

As can be seen, $z_{i+1}(x_0)$ actually only depends on $n_i(x_0)$ and not on the number of all the other haplotypes.

It then follows that $N_{i+1} = \sum_{x \in H_i} z_{i+1}(x)$ (the sum of the number of haplotypes in the $(i+1)$'th generation) conditionally on $\{n_i(x)\}_{x \in H_i}$ follows a Poisson$(\alpha_{i+1} N_i)$ distribution, and that

$$z_{i+1}(x_0) \mid \{n_i(x)\}_{x \in H_i}, N_{i+1} \sim \text{Binomial}\left(N_{i+1}, \frac{n_i(x_0)}{N_i}\right),$$

as expected, which is also true for the simple Fisher-Wright model in Equation (1).

2.3. *Mutation model*

As mentioned in the introduction, we assume a neutral (in the sense of no selection) single step mutation process on $\mathbb{Z}$. Instead of just one locus we extend it to r loci, where mutations on loci happen independently. We assume per locus and direction mutation rates. Let

$$Q = \{-1, 0, 1\}^r = \underbrace{\{-1, 0, 1\} \times \cdots \times \{-1, 0, 1\}}_{r \text{ factors}},$$

where $\times$ denotes the Cartesian product, be the lattice of possible mutations. Let

$$p_j(q) = \begin{cases} \delta_j & q = -1 \\ 1 - \delta_j - \omega_j & q = 0 \\ \omega_j & q = 1 \\ 0 & \text{else} \end{cases} \tag{4}$$

denote the mutation probabilities for the j'th locus and

$$p(q) = \prod_{j=1}^{r} p_j(q_j)$$

for a mutation configuration $q = (q_1, q_2, \ldots, q_r) \in Q$ from the fact that mutations are assumed to happen independently across loci.

Let

$$C_{i+1} = \bigcup_{\substack{q \in Q \\ x_1 \in H_i}} \{x_1 + q\}$$

be all possible candidate haplotypes for the $(i+1)$'th generation.

Our model with mutations is then

$$n_{i+1}(y_0) \mid \{n_i(x)\}_{x\in H_i} \sim \text{Poisson}\left(\alpha_{i+1}\sum_{q\in Q} p(q)n_i(y_0-q)\right) \quad \text{for all } y_0 \in C_{i+1}, \tag{5}$$

resulting in $N_{i+1} \mid N_i \sim \text{Poisson}(\alpha_{i+1}N_i)$ as assumed in Equation (2) because

$$\begin{aligned}\sum_{y_0\in C_{i+1}} \alpha_{i+1}\sum_{q\in Q} p(q)n_i(y_0-q) &= \alpha_{i+1}\sum_{q\in Q} p(q)\sum_{y_0\in C_{i+1}} n_i(y_0-q)\\ &= \alpha_{i+1}\sum_{q\in Q} p(q)N_i\\ &= \alpha_{i+1}N_i.\end{aligned}$$

Another way to formulate an equivalent model, which will be used in the implementation, is as follows. Let $m_{i+1}(x, x+q)$ denote the number of mutants mutating from x to $x+q$ in the transition from the i'th generation to the $(i+1)$'th generation and

$$M_{i+1}(x) = \{m_{i+1}(x, x+q)\}_{q\in Q}$$

the number of mutants for all possible configurations in Q.

Then assume that $\{M_{i+1}(x)\}_{x\in H_i}$ are conditionally independent given $\{z_{i+1}(x)\}_{x\in H_i}$, thus only the marginal distribution is to be specified. If we model this conditional marginal distribution as

$$M_{i+1}(x_0) \mid \{z_{i+1}(x)\}_{x\in H_i} \sim \text{Multinomial}\left(z_{i+1}(x_0), \{p(q)\}_{q\in Q}\right), \tag{6}$$

and set

$$n_{i+1}(x) = \sum_{q\in Q} m_{i+1}(x-q, x),$$

we get a model equivalent to the one specified in Equation (5).

2.4. *Absorbing state*

The model in Equation (5) (or the equivalent model in Equation (6)) has positive probability of dying out, because the Poisson distribution has probability mass in 0 for every parameter value. This means that population size 0 is an absorbing state. Also note that this absorbing state is independent of the mutation rate, as the population size is independent of the mutation rate.

3. Implementation

In this section, some implementation details are discussed. As already mentioned, the described model is implemented in the `R` (R Development Core Team, 2013) package `fwsim` (Andersen and Eriksen, 2012) using the `C` programming language. The package `fwsim` is released under the BSD license.

First some implementation details are explained and then a few examples are given.

3.1. *Haplotype container*

Each generation consists of a number of haplotypes, each with a count of the number of times it is present in the generation. These haplotypes are saved in a data container. This data container is a so-called k-d tree (Bentley, 1975) (this abbrivation stands for k dimensional tree), which is a generalisation of a binary search tree. Whereas binary search trees are for one dimensional points (numbers), k-d trees are for k dimensional points (vectors). Like binary search trees, the time complexity for insertion and searching in a k-d tree is $O(\log n)$ for a tree with n nodes.

For each generation, a new k-d tree is created and nodes inserted or updated as the haplotypes are evolved one at a time. A node in the tree contains both the point (haplotype) and additional information, which here is only a count (of the number of individuals having this particular haplotype).

The implementation of k-d trees is based on `http://code.google.com/kdtree` released under the BSD license, but has been heavily modified for example by changing some data structures and adding node searching and updating functionality.

3.2. *Mutation model*

In this section, the implementation of the mutation model defined in Section 2.3 is described.

The mutation model is implemented by dividing the number of children Equation (3) into categories depending on the number of times they mutate. There are $r+1$ categories, namely for $d = 0, 1, \ldots, r$ mutations on the r loci. Because this is the stepwise mutation model, only one mutation can happen per locus at a time.

As before, $z_{i+1}(x)$ is the number of children from haplotype $x \in H$. Let $z^d_{i+1}(x)$ be the number of children in the d'th category such that $z_{i+1}(x) = \sum_{d=0}^{r} z^d_{i+1}(x)$. If we assume that

$$z^d_{i+1}(x_0) \mid \{n_i(x)\}_{x \in H_i} \sim \text{Poisson}(\alpha_i \eta_d n_i(x_0)), \tag{7}$$

where η_d is the probability for d mutations with $\sum_d \eta_d = 1$, then Equation (3) still holds. Naturally, each of the $z^d_{i+1}(x)$ children have to choose their d mutations independently of the others.

To see the analogue between $m_{i+1}(x, x+q)$ and $z_{i+1}(x)$, first let

$$Q_d = \left\{ q \in Q \;\middle|\; \|q\|_1 = d \right\},$$

where $\|\cdot\|_1$ denotes the L^1 norm such that $\|q\|_1 = \|(q_1, q_2, \ldots, q_r)\|_1 = \sum_{j=1}^{r} |q_j|$. That is, Q_d is the mutation configurations resulting in precisely d mutations. Then

$$z^d_{i+1}(x) = \sum_{q \in Q_d} m_{i+1}(x, x+q).$$

First the probability of not mutating is treated. Let $\mu_j = \delta_j + \omega_j$ be the mutation rate for the j'th locus for $j = 1,2,\ldots,r$ with δ_j denoting the downwards mutation rate and ω_j denoting the upwards mutation rate. Then

$$\eta_0 = \prod_{j=1}^{r} (1 - \mu_j)$$

is the probability of not mutating.

Now the model of choosing the mutating loci is discussed. There are $\binom{r}{d}$ ways to choose the d loci that should mutate. Each of these loci configurations has 2^d possible mutation configurations (the size of the cartesian product $\{-1,1\}^d$). This means that there is a total of $2^d\binom{r}{d}$ possible ways to mutate d times. The probability for mutating to a specific haplotype is determined by the d locus specific upwards and downwards mutation rates.

For mutation category d, let

$$S_d = \left\{ s \subseteq \{1,2,\ldots,r\} \;\middle|\; |s| = d \right\}$$

be a so-called *simple table* with $\binom{r}{d}$ rows. Then the probability that it is exactly the loci $s \in S_d$ that should mutate, is

$$p(s) = \prod_{j \in s} \mu_j \prod_{j \in s^C} (1 - \mu_j),$$

where $s^C = \{1,2,\ldots,r,\} \setminus s$. Further, the probability of exactly d mutations is

$$\eta_d = \sum_{s \in S_d} p(s).$$

Hence, Equation (7) is now fully specified. To decide the direction of the mutations, let

$$E_d = \left\{ (s,q) \mid s \in S_d, q : s \to \{-1,1\}^d \right\}$$

be a so-called *extended table* with $2^d\binom{r}{d}$ rows. The function q maps a locus to a mutation direction. Then each row $e = (s,q) \in E_d$ and has probability

$$p(e) = \prod_{j \in s} p_j(q(j)) \prod_{j \in s^C} (1 - p_j(q(j))),$$

where $p_j(q(j))$ is defined in Equation (4). We still have that the sum of the rows in the extended table is η_d.

Then for generation i, haplotype x, and mutation category d, we assume that

$$\{m_{i+1}(x_0, x_0 + q)\}_{q \in Q_d} \mid \{n_i(x)\}_{x \in H_i} \sim \text{Multinomial}\left(z_{i+1}^d(x_0), \left\{ \frac{p(e)}{\sum_{e \in E_d} p(e)} \right\}_{e \in E_d} \right).$$

Both the simple and extended table for mutation category $d = 1,2,\ldots,r$ ($d = 0$ does not require this step) are created before the actual simulation starts as

the probabilities are constant during the evolution. They are constant because the mutation rates are assumed constant. This is what is done in the `fwsim` package for all mutation categories, although this may be changed in future releases if the following theoretical limitations turn out to occur in practise, too.

Note that $2^d\binom{r}{d}$, the size of the extended table, is exponentially growing and may become really large for even relatively small r and that the corresponding extended tables take some time to generate. For example, for $r = 16$ and $d = 11$ the size of the extended table is $8,945,664$ (the maximal for that choice of r), however, it is still possible to be created and used for simulation. Once the tables are created, the simulations run rather smoothly because they are just stored in memory.

On the other hand, the mutation rate would normally be so low that mutations in the categories for even small d may rarely or never happen depending on the population size, which means that these mutation categories are probably better delt with manually as follows. Recall that η_d only depends on the simple table, which is small compared to the extended table – namely a factor of 2^d smaller – and so the simple table can still be calculated to a rather large r. When the simple tables are generated, then draw n from $\text{Poisson}(\alpha_{i+1}\eta_d n_i(x))$ and mutate each of the haplotypes manually one at a time by choosing the d loci and their directions randomly according to their probabilities.

4. Computation time

The simulation method described above is developed with efficiency in mind. To illustrate that efficiency is achieved, the computation time for different parameters have been investigated using a laptop with a 2.40GHz Intel(R) Core(TM) i5 CPU (model M 520). For these computations, `fwsim` (Andersen and Eriksen, 2012) version 0.2-5 was used.

In Figure 1, the absolute computation time for simulating a population with a varying number of loci is shown. In Figure 1, the computation time for simulating a population with a varying initial population size is shown. Both figures show that the algorithm is quite fast.

In Table 1, the computation time using `fwsim` compared to a naïve implementation (focusing on individuals rather than haplotypes) of simulating under a Fisher-Wright model is shown. As seen, `fwsim` is magnitudes faster than a naïve implementation: On average, `fwsim` is almost 2,000 times faster when simulating a population with an initial size of 5,000, no expected growth (by using the growth parameter $\alpha = 1$), and a mutation rate of 0.003 in 100 generations than the naïve implementation (focusing on individuals rather than haplotypes). Further, the memory consumption is smaller for `fwsim` as it uses haplotypes instead of individuals, which means that it is possible to simulate much larger populations than with a naïve implementation.

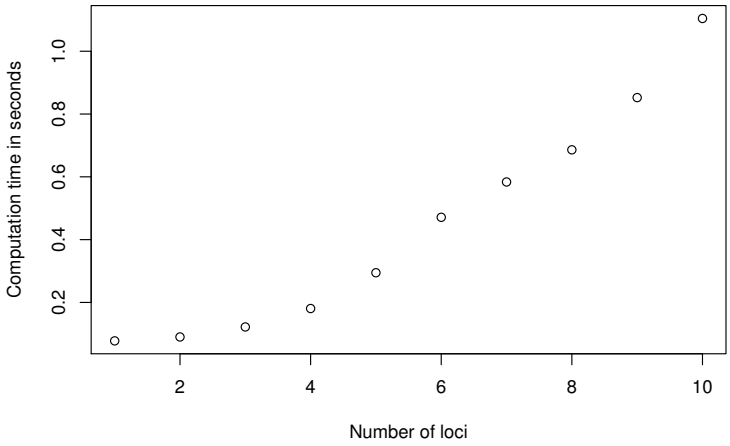


Figure 1. The computation time depending on the number of loci. The initial population size is set to 10,000, the number of generations to 500, the mutation rate to 0.003, the growth parameter to 1 (meaning constant expected population size). The computation time for each number of loci is the median computation time of 10 simulations.

5. Examples

In this section, some examples are presented. Please refer to `?fwsim` in `R` for more information about usage of the package `fwsim`. These examples were made using version 0.2-5 of `fwsim` (Andersen and Eriksen, 2012).

5.1. *Simple usage*

Lauching an `R` session and typing the code below will show a short example of the model implemented in the package `fwsim` (`k` is the number of individuals in the initial population, `g` is the number of generations to evolve, `r` number of loci, `mu` mutation rate per loci, `alpha` is the population size growth rate and `trace` is whether to display trace information):

```
> library(fwsim)
> set.seed(1)
> pop <- fwsim(k = 10000, g = 1000, r = 3, mu = 0.003,
>   alpha = 1.001, trace = TRUE)
```

To obtain a contingency table of the first two loci, use the following:

```
> sum(pop$haplotypes$N)
> [1] 27672
> xtabs(N ~ Locus1 + Locus2, pop$haplotypes)
      Locus2
Locus1   -5   -4   -3   -2   -1    0    1    2    3    4    5    6
    -6    0    0    0    0    5    2    0    0    0    0    0    0
```

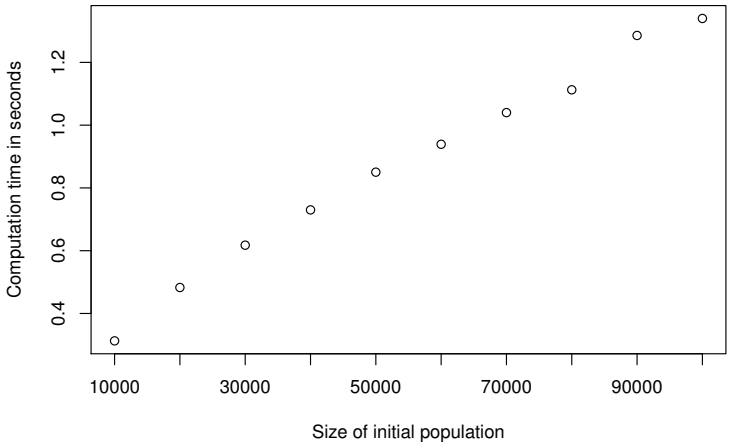


Figure 2. The computation time depending on the initial population size. The number of loci is set to 5, the number of generations to 500, the mutation rate to 0.003, the growth parameter to 1 (meaning constant expected population size). The computation time for each number of loci is the median computation time of 10 simulations.

```
     -5    0    0    0   10   85   75   10    0    2    0    0    0
     -4    0    0    0   23   46  301   37  118    1    0    0    0
     -3    0    0   15  394  591  474  266  122  110    5    1    0
     -2    0   11  144  723  717 1302  542  526   17    9   26    0
     -1    0  108  148 1018 1048 1816 1039  453  517  197  138  101
     0     1   30  347  879 1713  901 1038  509  448  184   27   11
     1    34  198  647  552  324  715  810  421  300   90   11    0
     2     0   63   37  659  349  492  314  306  105   10    0    0
     3     0   73  420  540  290   50   30  160    0    0    0    0
     4     0   20   58   94   63    4   41    2    0    0    0    0
     5     0    0    9    0    0    0    0    0    0    0    0    0
```

This table is plotted in Figure 3. A slight drift from the initial $(0,0)$ has occured.

We can also see the 10 most frequent haplotypes compared to the initial $(0,0,0)$ haplotype:

```
> pop$haplotypes[order(pop$haplotypes$N, decreasing = TRUE)[1:10], ]
    Locus1 Locus2 Locus3   N
279     -1      0      0 665
105     -1     -2     -2 539
270     -1      0     -1 517
269     -2      0     -1 509
173      0     -1     -1 482
160      0     -1     -2 423
179      0     -1      0 423
341     -1      1     -1 385
274     -2      0      0 378
241     -1      0     -2 358
> pop$haplotypes[which(apply(apply(pop$haplotypes[, 1:3], 1, abs),
>   2, sum) == 0), ]
```

Computation time speed-up			
k	g	μ	Speed-up
1,000	100	0.001	145.9
1,000	100	0.003	127.2
1,000	200	0.001	307.9
1,000	200	0.003	372.5
5,000	100	0.001	2,972.1
5,000	100	0.003	1,957.0
5,000	200	0.001	6,848.4
5,000	200	0.003	4,887.1

Table 1. A comparison of the computation time for `fwsim` and a naïve implementation (focusing on individuals rather than haplotypes). A growth parameter $\alpha = 1$ is used meaning no expected population growth. k is the initial population size, g is the number of generations to evolve, and μ is the mutation rate. 10 replications for each parameter combination (corresponding to a row in the table) were performed. The speed-up column is the computation time for the naïve implementation divided by the computation time for `fwsim`. This means that `fwsim` on average is roughly 2,000 times faster to simulate a population with an initial size of 5,000 and a mutation rate of 0.003 in 100 generations than the naïve implementation.

```
    Locus1 Locus2 Locus3   N
280      0      0      0 255
```

In Figure 4, the actual population sizes are compared to expected population sizes. This figure was made with following code:

```
> plot(pop$sizes, type = "l", xlab = "Generation",
>   ylab = "Population size", lty = 1)
> lines(pop$expected.sizes, lty = 2)
> legend("topleft", legend = c("Actual", "Expected"), lty = 1:2)
```

5.2. *Genetic drift of alleles*

To illustrate how genetic drift in terms of changed allele frequencies occurs, the allele frequencies after a different number of generations are recorded. The `fwsim` package also has the possibility of saving the intermediate populations, which is used to show how allele frequencies change during the evolution. Thus, genetic drift can be investigated as follows (`k` is the number of individuals in the initial population, `alim` is the limit of which alleles to plot and `gs` is which generations to sample allele frequencies from):

```
> library(fwsim)
> set.seed(1)
> alim <- 2
> k <- 100000000
> g <- 10000
> gs <- seq(100, g - 1, by = 100)
> pop <- fwsim(g = g, k = k, r = 1, alpha = 1,
>   mu = 0.003, gs = gs, trace = FALSE)
> interhapfreq <- lapply(pop$intermediate.haplotypes[gs], function(hap) {
>   tab <- prop.table(xtabs(N ~ Locus1, hap))
```

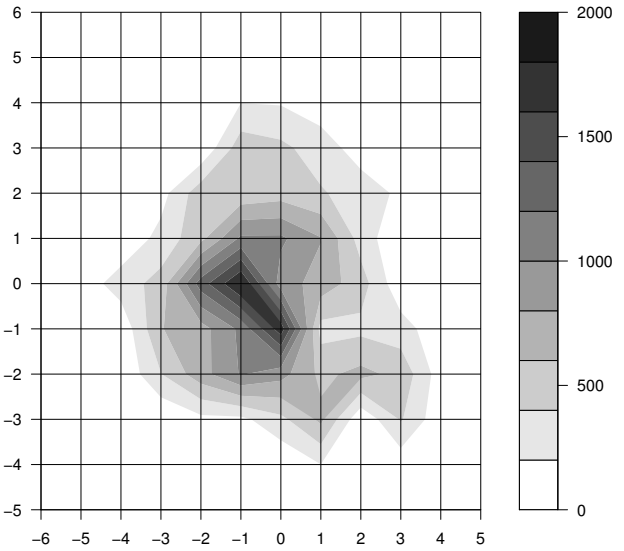


Figure 3. A contour plot of the contingency table of the first two loci. A slight drift from the initial $(0,0)$ has occured.

```
>   as.vector(tab[which(abs(as.numeric(names(tab))) <= alim)])
> })
> freq <- data.frame(do.call("rbind", interhapfreq))
> colnames(freq) <- (-alim):alim
> plot(gs, freq[, alim+1], type = "l",
>   xlab = "Number of generations",
>   ylab = "Frequency", ylim = range(freq))
> for (a in 1:alim) {
>   i1 <- (alim+1)-a
>   i2 <- (alim+1)+a
>   lines(gs, freq[, i1], type = "l", lty = a + 1)
>   lines(gs, freq[, i2], type = "l", lty = a + 1)
> }
> others <- 1-apply(freq, 1, sum)
> lines(gs, others, type = "l", lty = alim + 2)
> legend("topright",
>   legend = c(paste("Allele", c(0, paste("+/-", 1:alim))),
>     "Other alleles"),
>   lty = 1:(alim+2))
```

Note that we only simulate one locus and set the population size quite large to get the asymptotic behaviour. The resulting plot can be seen in Figure 5.

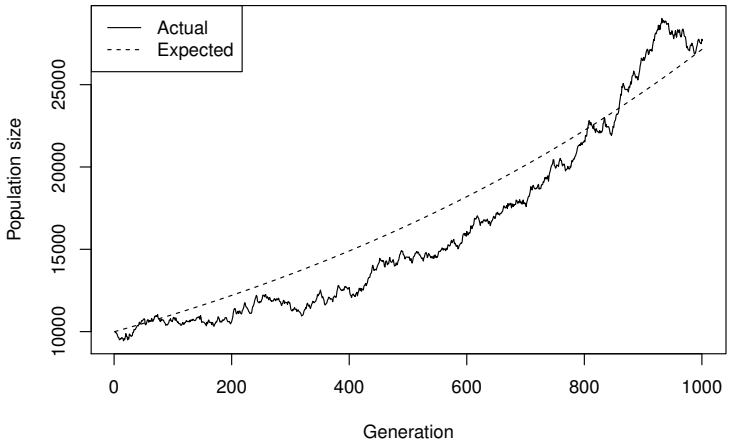


Figure 4. The actual population sizes compared to expected population sizes

5.3. *Genetic drift of alleles depending on mutation rate*

To illustrate how genetic drift in terms of changed allele frequencies for the 0 allele occurs depending on the mutation rate, the allele frequencies after a different number of generations are recorded for populations with different mutation rates. Thus, genetic drift depending on mutation rate may be investigated as follows (`k` is the number of individuals in the initial population and `gs` is which generations to sample allele frequencies from):

```
> library("fwsim")
> mus <- c(0.001, 0.002, 0.003)
> k <- 100000000
> g <- 10000
> gs <- seq(100, g - 1, by = 100)
> set.seed(1)
> freqs <- lapply(mus, function(mu) {
>   pop <- fwsim(g = g, k = k, r = 1, alpha = 1, mu = mu, save.gs = gs,
>     trace = FALSE)
>   sapply(pop$intermediate.haplotypes[gs],
>     function(hap) hap$N[which(hap[, 1] == 0)] / sum(hap))
> })
> plot(gs, freqs[[1]], type = "l",
>   xlab = "Number of generations", ylab = "Frequency for allele 0",
>   ylim = range(unlist(lapply(freqs, range))), lty = 1)
> for (i in 2:length(mus)) lines(gs, freqs[[i]], type = "l", lty = i)
> legend("topright", legend = paste("mu = ", mus, sep = ""),
>   lty = 1:length(mus))
```

Note that we only simulate one locus and set the population size quite large to get the asymptotic behaviour. The resulting plot can be seen in Figure 6.

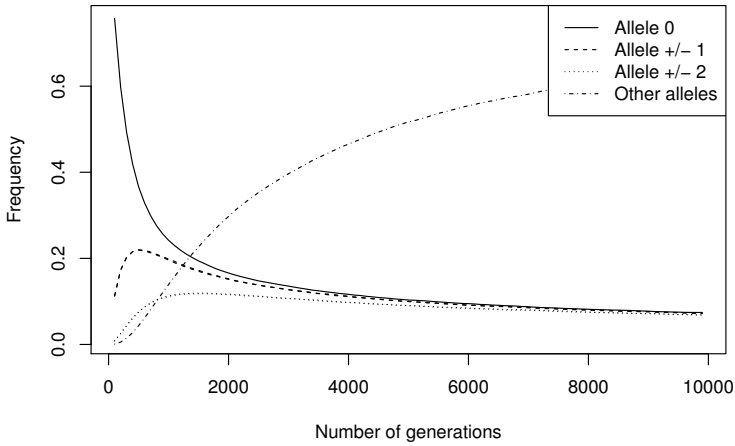


Figure 5. Simulated genetic drift using an initial population of size 100,000,000, a growth of 1 (meaning no expected growth), and a mutation rate of 0.003.

Acknowledgement

The authors would like to thank Torben Tvedebrink, PhD; Søren Højsgaard, PhD; and Lisbeth Grubbe Nielsen, all Aalborg University, Denmark, for helping us improving the manuscript.

6. Bibliography

Andersen, M. M. and Eriksen, P. S. (2012) *fwsim: Fisher-Wright Population Simulation*. URL `http://CRAN.R-project.org/package=fwsim`. R package version 0.2-5. 62, 66, 69, 70

Bentley, J. L. (1975) Multidimensional Binary Search Trees Used for Associative Searching. *Communications of the ACM*, **18**, 509–517. 67

Butler, J. M. (2005) *Forensic DNA Typing: Biology, Technology, and Genetics of STR Markers*. Academic Press, 2nd edn. 62

Caliebe, A., Jochens, A., Krawczak, M. and Rösler, U. (2010) A Markov Chain Description of the Stepwise Mutation Model: Local and Global Behaviour of the Allele Process. *Journal of Theoretical Biology*, **266**, 336–342. 62

Ewens, W. J. (2004) *Mathematical Population Genetics*. Springer-Verlag. 62

Fisher, R. A. (1922) On the Dominance Ratio. *Proc. Roy. Soc. Edin.*, **42**, 321–341. 62

Fisher, R. A. (1930) *The Genetical Theory of Natural Selection*. Oxford: Clarendon Press. 62

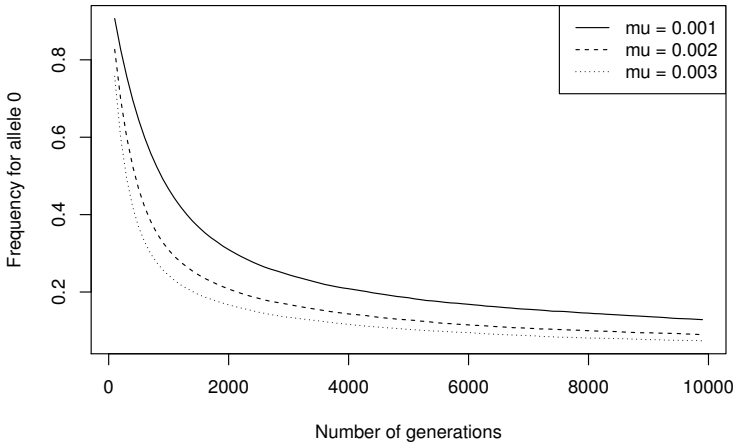


Figure 6. Simulated genetic drift using a population size of 100,000,000 and a growth of 1 (meaning no expected growth).

Fisher, R. A. (1958) *The Genetical Theory of Natural Selection*. New York: Dover, 2nd revised edn. 62

Hein, J., Schierup, M. H. and Wiuf, C. (2005) *Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory*. Oxford University Press. 62

Hudson, R. R. (2001) Generating Samples Under a Wright–Fisher Neutral Model of Genetic Variation. *Bioinformatics*, **18**. 62

Kingman, J. F. C. (1982) The Coalescent. *Stochastic Processes and their Applications*, **13**, 235–248. 62

Ohta, T. and Kimura, M. (1973) A Model of Mutation Appropriate to Estimate the Number of Electrophoretically Detectable Alleles in a Finite Population. *Genet. Res.*, **22**, 201–204. 62

R Development Core Team (2013) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org`. ISBN 3-900051-07-0. 62, 66

Wright, S. (1931) Evolution in Mendelian populations. *Genetics*, **16**, 97–159. 62

Paper VI

The discrete Laplace exponential family and estimation of Y-STR haplotype frequencies

Author list Mikkel Meyer Andersen, *Aalborg University, Denmark*
Poul Svante Eriksen, *Aalborg University, Denmark*
Niels Morling, *University of Copenhagen, Denmark*

Summary Estimating haplotype frequencies is important in e.g. forensic genetics, where the frequencies are needed to calculate the likelihood ratio for the evidential weight of a DNA profile found at a crime scene. Estimation is naturally based on a population model, motivating the investigation of the Fisher-Wright model of evolution for haploid lineage DNA markers.
An exponential family (a class of probability distributions that is well understood in probability theory such that inference is easily made by using existing software) called the 'discrete Laplace distribution' is described. We illustrate how well the discrete Laplace distribution approximates a more complicated distribution that arises by investigating the well-known population genetic Fisher-Wright model of evolution by a single-step mutation process.
It was shown how the discrete Laplace distribution can be used to estimate haplotype frequencies for haploid lineage DNA markers (such as Y-chromosomal short tandem repeats), which in turn can be used to assess the evidential weight of a DNA profile found at a crime scene. This was done by making inference in a mixture of multivariate, marginally independent, discrete Laplace distributions using the EM algorithm to estimate the probabilities of membership of a set of unobserved subpopulations. The discrete Laplace distribution can be used to estimate haplotype frequencies with lower prediction error than other existing estimators. Furthermore, the calculations could be performed on a normal computer.
This method was implemented in the freely available open source software `R` that is supported on Linux, MacOS and MS Windows.

Publication info This paper was published as:

> Andersen MM, Eriksen PS, Morling N (2013). *The discrete Laplace exponential family and estimation of Y-STR haplotype frequencies.* Journal of Theoretical Biology; **329**:39-51.

The version here is the journal version with correction of moments of certain random variables (e.g. $\mathbf{E}[D]$ changed to $\mathbf{E}[|D|]$) that are also published as a corrigendum, minor typographical corrections and mentioning newer versions of the software (the `R` packages `disclap` and `disclapmix`).

1. Introduction

The use of haploid lineage DNA markers such as Y-chromosomal short tandem repeats (Y-STRs) or mitochondrial DNA (mtDNA) polymorphisms have important advantages in certain types of forensic genetic casework (Gill *et al.*, 1985; Sibille *et al.*, 2002; Roewer, 2009). If e.g. only a small amount of male DNA is found in combination with a large amount of female DNA, Y-STR typing may be very valuable. If e.g. only a hair shaft is found, mtDNA typing may assist in solving the case. We focus on Y-STR in this paper and note that many of the properties of Y-STR are true for mtDNA as well, because they are both lineage markers.

A very important task in forensic genetics is to evaluate the evidential weight of the evidence by means of likelihood principles (Evett and Weir, 1998; Gill *et al.*, 2001). The likelihood ratio used is

$$LR = \frac{P(E|H_p)}{P(E|H_d)},$$

where H_p is the prosecutor's hypothesis (e.g. 'The suspect is the donor of the genetic data') and H_d is the defense attorney's hypothesis (e.g. 'The suspect is not connected to the crime').

In most single doner cases where it is assumed that errors do not happen, it is often assumed that $P(E|H_p) = 1$. Then $P(E|H_d)$ is called the 'match probability' and is often interpreted as the probability that an individual drawn randomly from the population has the same DNA profile as the trace found at a crime scene. Note, that if we knew the haplotypes of the entire population, the population frequency of the haplotype in question would be the match probability (in an idealised population without e.g. population structure). Thus, assuming a simple population model, the match probability is the haplotype frequency of the haplotype found at the crime scene.

Due to the lack of recombination, there is statistical dependence between loci, making calculations of match probabilities of lineage markers more challenging than that of autosomal markers (Buckleton *et al.*, 2011). Naïve counts/estimates of match probabilities in a reference database of size n and a haplotype observed x times like x/n, $(x+1)/(n+1)$ or similar seem to be rather conservative and not generally satisfactory (Brenner, 2010; Buckleton *et al.*, 2011). The method of Roewer *et al.* (2000); Krawczak (2001); Willuweit *et al.* (2011) takes the evolutionary aspect of Y-STRs into consideration (see http://www.yhrd.org). Unfortunately, it seems to have some draw backs as indicated by e.g. Andersen (2010). Brenner (2010) suggested a method that takes the rarity of Y-STR haplotypes into consideration. In particular, when considering Y-STR haplotypes comprising a large number of genetic loci, the proportion of haplotypes observed only once – singletons – will be high. Brenner (2010) suggested to adjust/correct the match probability of singletons with a factor, κ, that reflects the ratio between singletons and non-singletons (Robbins, 1968). The κ correction method estimates the match probability by $(1-\kappa)/(n+1)$, where $\kappa = (\alpha+1)/(n+1)$ and α denotes the total number of singletons in the reference database. This method was discussed by Buckleton *et al.* (2011) and Andersen *et al.* (2013a).

We have developed a model based on assumptions of primarily neutral, single-step mutations of STRs (Ohta and Kimura, 1973) that are following the Fisher-Wright model of evolution (Fisher, 1922, 1930, 1958; Wright, 1931; Ewens, 2004). Caliebe *et al.* (2010) discussed certain properties of a Fisher-Wright model with neutral single-step mutations. They found the distribution of a quantity that they refer to as the normalised allele process. In this paper, we describe this process and suggest an approximation to its distribution that turn out to be an exponential family called the 'discrete Laplace distribution' due to its similarities to the Laplace distribution of real numbers. This distribution has been described by Inusah and Kozubowski (2006), although they do not note that it is actually an exponential family.

Finally, examples of the use of the discrete Laplace distribution for the estimation of haplotype frequencies for Y-STRs are presented and compared to the results obtained with other methods. The discrete Laplace distribution was used as a family function in a generalized linear model (GLM). The EM algorithm (Dempster *et al.*, 1977) was used to estimate the probability of membership of a set of unobserved subpopulations. The calculations could be performed on a normal computer: Haplotype frequencies of a database with 1,000 Y-STR haplotypes consisting of 7 loci could be estimated in around 0.025 seconds assuming 1 subpopulation, in around 0.6 seconds assuming 2 subpopulations and in around 2.9 seconds assuming 5 subpopulations using a Lenovo T410s laptop with 6 GB RAM and an Intel® Core™ i5 CPU model M520 running at 2.40GHz.

Thus, this paper consists of two parts: (1) an introduction to an exponential family – the discrete Laplace distribution — and (2) an analysis of the application of it in the analyses of lineage markers in population and forensic genetics. Three `R` (R Development Core Team, 2013) packages 'fwsim' (Andersen and Eriksen, 2012b) (submitted, see Andersen and Eriksen (2012a) for a preprint), 'disclap' (Andersen and Eriksen, 2013a), and 'disclapmix' (Andersen and Eriksen, 2013b) were produced. 'fwsim' (`http://cran.r-project.org/package=fwsim`) simulates populations under the Fisher-Wright model, 'disclap' (`http://cran.r-project.org/package=disclap`) implements the exponential family and 'disclapmix' (`http://cran.r-project.org/package=disclapmix`) uses the EM algorithm (Dempster *et al.*, 1977) to perform inference for a mixture of distributions. Please, refer to Andersen *et al.* (2013b) for an introduction on how to use these software packages.

2. Discrete Laplace distribution

In this section, the normalised allele process of Caliebe *et al.* (2010) is described. The discrete Laplace distribution (or double geometric distribution) is introduced as a simple probability distribution. An approximation of the distribution of the normalised allele process in terms of the discrete Laplace distribution is discussed and introduced as an exponential family.

2.1. *Motivation*

Let N be a constant population size and let $X_g(i) \in \mathbb{Z}$ denote the STR allele (number of repeats) of the i'th individual in the g'th generation. Thus, it is assumed that alleles are integers. This immediately rules out 'null alleles' (typically a SNP in the primer binding regions of around the Y-STR), intermediate alleles and duplications (Butler, 2005; Budowle *et al.*, 2008). This is a well-known limitation to mathematical STR models that for example coalescent theory also suffers from (Hein *et al.*, 2005; Andersen *et al.*, 2013a). The normalised allele process is

$$V_g(i) := X_g(i) - X_g(N) \quad \text{for } i \neq N. \tag{1}$$

The normalised allele process has a mean value of zero. It is a positively recurrent, irreducible, and aperiodic Markov chain that converges exponentially fast to the unique unimodal invariant distribution (Caliebe *et al.*, 2010).

Motivated by the results by Caliebe *et al.* (2010) – especially the simulation results shown in (Caliebe *et al.*, 2010, Figure 1) for certain choices of N, mutation rate, and number of generations – the distribution of the normalised allele process can be approximated by a distribution similar to that of the geometric distribution, but with $\mathbb{Z}$ as support instead of just $\{0,1,\ldots\}$. We refer to this distribution as the 'discrete Laplace distribution'.

2.2. *A simple probability distribution*

The random variable D follows a discrete Laplace distribution with parameter $0 < p < 1$ if its probability mass function is such that $P(D = d) \propto p^{|d|}$.

The normalisation constant is found by considering the double geometric series

$$\sum_{d \in \mathbb{Z}} p^{|d|} = \frac{1+p}{1-p},$$

such that

$$P(D = d) = \left(\frac{1-p}{1+p}\right) p^{|d|}$$

for $0 < p < 1$ and $d \in \mathbb{Z}$. Later, in Section 2.5, it is shown that

$$\mathbf{E}[|D|] = \frac{2p}{1-p^2}. \tag{2}$$

2.3. *Approximating the normalised allele process*

The interesting quantity is the distribution of Equation (1), where Caliebe *et al.* (2010) refers to the probability mass function as η, such that

$$\eta_g(d) = P(V_g(i) = d) \tag{3}$$

for $d \in \mathbb{Z}$. Let $Z_j(i) \in \{-1,0,1\}$ be the mutation event preceding the inheritance of the i'th individual in the j'th generation. For easier notation, first let

$$Q_j(i) = Z_j(i) - Z_j(N) + 2.$$

If μ is the mutation probability, then

$$q(d) := P(Q_j(i) = d) = \begin{cases} \mu^2/4 & \text{if } d = 0, \\ \mu - \mu^2 & \text{if } d = 1, \\ 1 - 2\mu + 3\mu^2/2 & \text{if } d = 2, \\ \mu - \mu^2 & \text{if } d = 3, \\ \mu^2/4 & \text{if } d = 4, \\ 0 & \text{else.} \end{cases}$$

Thus, $q(d) = r(d-2)$ in the notation of Caliebe *et al.* (2010) (but as we use r as the number of loci, this function will not be used any further). Let

$$\gamma_g(d) = \eta_g(d + 2g). \tag{4}$$

Two expressions of Equation (3) and Equation (4) were derived in (Caliebe *et al.*, 2010). The first is a recurrence relation (Caliebe *et al.*, 2010, Lemma 8). The second is a sum of probability mass function convolutions (Caliebe *et al.*, 2010, Theorem 13), which reformulated in terms of γ_g instead of η_g can be expressed as

$$\gamma_g = \frac{1}{N} q * \left(\sum_{i=0}^{g-2} \left[\frac{N-1}{N} \right]^i q^i \right) + \left(\frac{N-1}{N} \right)^{g-1} q^g$$

for $g \in \{2,3,\ldots\}$, where $*$ means the convolution and $q^i = q^{i-1} * q$ means the i'th convolution of q.

Using the recurrence relation, Caliebe *et al.* (2010) plotted this density, which we will compare to an approximation by the discrete Laplace distribution. First, an alternative way of calculating $\eta_g(d)$, and thus $\gamma_g(d)$ numerically, will be described. This method exploits how to do convolutions quickly using a discrete Fourier transformation (Cooley *et al.*, 1969; Brigham, 1988).

By definition

$$\mathbf{E}\left(\theta^{Q_j}\right) = \sum_{d=0}^{4} P(Q_j = d)\theta^d = \sum_{d=0}^{4} q(d)\theta^d$$

for some $\theta \in \mathbb{C}$, which results in

$$\mathbf{E}\left(\theta^{\sum_{j=1}^{g} Q_j}\right) = \left(\sum_{d=0}^{4} q(d)\theta^d \right)^g = \sum_{d=0}^{4g} q^g(d)\theta^d$$

due to independence.

Let

$$\theta_a = e^{-2\pi \mathbf{i} a/(4g+1)}$$

for $a = 0, 1, \ldots, 4g$, where $\mathbf{i}$ is the imaginary unit satisfying $\mathbf{i}^2 = -1$, and define

$$X_a = \left(\sum_{d=0}^{4} q(d) \theta_a^d \right)^g .$$

Then by Fourier inversion,

$$q^g(d) = \sum_{a=0}^{4g} X_a e^{2\pi \mathbf{i} d a/(4g+1)}.$$

Hence, $q^g(d)$ can be found by a fast Fourier transformation (FFT) algorithm, e.g. by using the `fft` function in `R` (R Development Core Team, 2013). When the convolutions are calculated, the value of $\eta_g(d)$ is also quickly calculated.

We suggest that the discrete Laplace distribution approximates the distribution of the normalised allele process, $\eta_g(d) = P(V_g(i) = d)$, in (Caliebe *et al.*, 2010). We compared the figures (Caliebe *et al.*, 2010, Figure 1 and Figure 2), see Figure 1 and Figure 2 with the approximating discrete Laplace distribution. For each set of parameters, the corresponding parameter, p, of the discrete Laplace distribution was found by calculating the mean,

$$\mu = \mathbf{E}[|V_g(i)|] = 2 \sum_{d=1}^{2g} d \eta_g(d),$$

and solving Equation (2) for p to obtain this parameter.

In Figure 3, a probably more realistic mutation rate for Y-STR of $\mu = 0.003$ (Ballantyne *et al.*, 2010) was used.

2.4. *Approximation properties*

To investigate the approximation properties, the Kullback-Leibler distance (Kullback and Leibler, 1951; Kullback, 1959) between the exact distribution, η_g, given in Equation (3) (or γ_g given in Equation (4)) and the discrete Laplace distribution was calculated. Assume that D is distributed according to a discrete Laplace distribution and let $f(d) = P(D = d)$. Let

$$\mathrm{KL}(\eta_g, f) = \sum_{d \in \mathbb{Z}} \eta_g(d) \log\left(\frac{\eta_g(d)}{f(d)} \right) = \sum_{d=-g}^{g} \eta_g(d) \log\left(\frac{\eta_g(d)}{f(d)} \right)$$

as $0 \log 0 = 0$.

The Kullback-Leibler distances for different mutation rates, number of generations and number of individuals are shown in Figure 4. As seen, the error increases with the mutation rate (to some asymptotic value, it seems). Given a fixed number of generations, the error also increases with the number of individuals. On the other hand, given a fixed number of individuals, there are some points where the lines cross and the number of generations causing the largest error depend on the mutation rate.

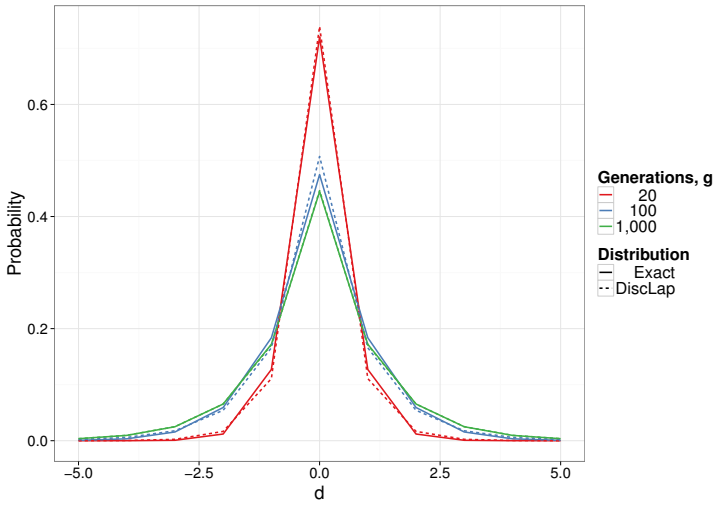


Figure 1. Exact probability, $\eta_g(d) = P(V_g(i) = d)$, for various values of generations, g, with population size $N = 100$ and mutation rate $\mu = 0.01$ and the corresponding approximation by the discrete Laplace distribution (DiscLap).

2.5. *An exponential family*

Assume that the signed allele distance, $d \in \mathbb{Z}$, from an ancestor is distributed according to the probability mass function given by

$$f(d;p) = \left(\frac{1-p}{1+p}\right) p^{|d|}, \tag{5}$$

where $0 < p < 1$ is the parameter of the model and $(1-p)/(1+p)$ is the normalising constant. A reparameterisation with

$$\theta = \log p,$$

such that $\theta < 0$ shows that this is an exponential family, because

$$f(d;\theta) = \exp\left(\log\left(\frac{1-e^{\theta}}{1+e^{\theta}}\right) + \theta|d|\right) = \exp\left(\theta|d| - A(\theta)\right)$$

with

$$A(\theta) = \log\left(\frac{1+e^{\theta}}{1-e^{\theta}}\right).$$

The probability mass function `ddisclap`, cumulative distribution function `pdisclap`, random deviates generation function `rdisclap` and family object generation function `DiscreteLaplace` for this exponential family were implemented in the `R` (R Development Core Team, 2013) package `disclap` (Andersen and Eriksen, 2013a).

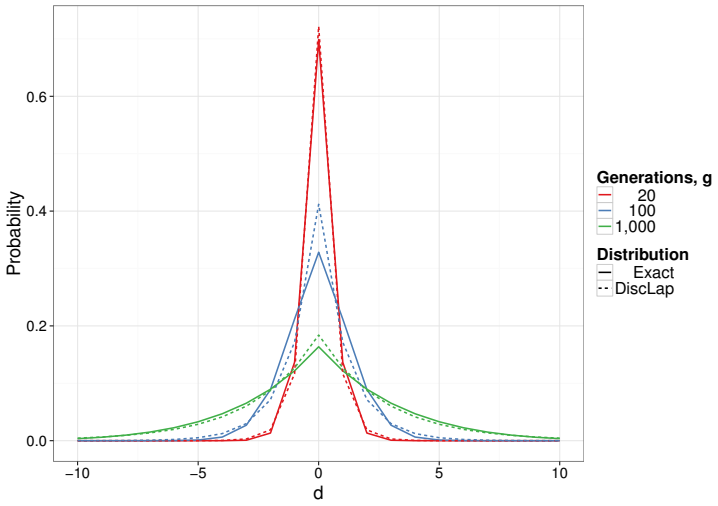


Figure 2. Exact probability, $\eta_g(d) = P(V_g(i) = d)$, for various values of generations, g, with population size $N = 1{,}000$ and mutation rate $\mu = 0.01$ and the corresponding approximation by the discrete Laplace distribution (DiscLap).

Cumulants

We now proceed with the cumulants to easily obtain the mean value and the variance function of the distribution. Let D have the probability mass function, $f(d; p)$, as defined in Equation (5). Then,

$$\mu = \mathbf{E}[|D|] = \frac{\partial A(\theta)}{\partial \theta} = \frac{\partial p}{\partial \theta}\frac{\partial A}{\partial p} = \frac{2p}{1-p^2}.$$

Furthermore, we obtain the variance function as

$$v(\mu) = \mathbf{Var}[|D|] = \frac{\partial \mu}{\partial \theta} = \frac{\partial p}{\partial \theta}\frac{\partial \mu}{\partial p} = \mu\left(\frac{1+p^2}{1-p^2}\right).$$

Solving $\mu = 2p/(1-p^2)$ for p, yields

$$p = \mu^{-1}(\sqrt{\mu^2+1} - 1), \tag{6}$$

making it possible to obtain the variance function as a function of the mean, i.e.

$$v(\mu) = \mu\sqrt{1+\mu^2}.$$

For practical purposes, in the implementation of the generalized linear model family in R (R Development Core Team, 2013), it is useful to also have the probability mass function as a function of the mean, which is obtained by

$$f(d; p) = \left(\frac{\mu - \sqrt{1+\mu^2}+1}{\mu+\sqrt{1+\mu^2}-1}\right) \times \left(\sqrt{1+\mu^2}-1\right)^{|d|} \mu^{-|d|}.$$

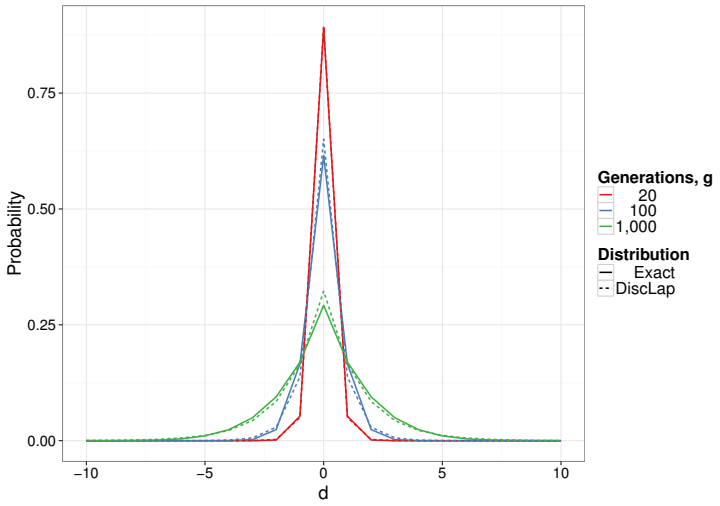


Figure 3. Exact probability, $\eta_g(d) = P(V_g(i) = d)$, for various values of generations, g, with population size $N = 1{,}000$ and mutation rate $\mu = 0.003$ and the corresponding approximation by the discrete Laplace distribution (DiscLap).

Link function

The canonical link function, g, is found as $g(\mu) = \theta = \log p$, which is equivalent to

$$\theta = g(\mu) = \log\left(\frac{\sqrt{1+\mu^2}-1}{\mu}\right).$$

Deviance

Let

$$L(p;d) = f(d;p) = \left(\frac{1-p}{1+p}\right)p^{|d|}.$$

From Equation (6),

$$p = p(\mu) = \mu^{-1}(\sqrt{\mu^2+1}-1),$$

yielding

$$\begin{aligned} l(\mu;d) &= \log L(p(\mu);d) \\ &= \log\left(\frac{1-p(\mu)}{1+p(\mu)}\right) + |d|\log\big(p(\mu)\big) \\ &= \log\left(\frac{1-\mu^{-1}(\sqrt{\mu^2+1}-1)}{1+\mu^{-1}(\sqrt{\mu^2+1}-1)}\right) + \end{aligned}$$

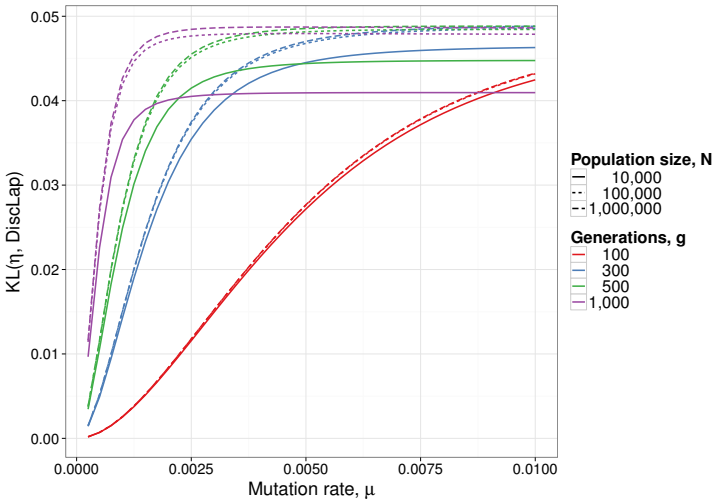


Figure 4. Kullback-Leibler distance between the exact distribution, η_g, and the approximating discrete Laplace distribution.

$$|d|\log\left(\mu^{-1}(\sqrt{\mu^2+1}-1)\right).$$

The deviance for one observation, d, is

$$D_1(d,p) = -2\log\frac{L(p(\mu);d)}{L(p(d);d)} = -2(l(\mu;d)-l(d;d)) = 2(l(d;d)-l(\mu;d)).$$

In the special case, where $d=0$, we use L'Hôpital's rule (also called Bernoulli's rule) to find the limit using the derivatives of the numerator and denominator and obtain

$$\lim_{d\to 0}\frac{\sqrt{d^2+1}-1}{d} = \lim_{d\to 0}\frac{d\frac{1}{\sqrt{d^2+1}}}{1} = \lim_{d\to 0}\frac{1}{\sqrt{1+\frac{1}{d^2}}} = 0$$

such that for $d=0$,

$$l(d;0) = \log 1 + 0\log 0 - \log 1 = 0$$

and

$$l(\mu;0) = \log\left(\frac{1-\mu^{-1}(\sqrt{\mu^2+1}-1)}{1+\mu^{-1}(\sqrt{\mu^2+1}-1)}\right).$$

To summarise,

$$D_1(d,p) = \begin{cases} 2\log\left(\frac{1+\mu^{-1}(\sqrt{\mu^2+1}-1)}{1-\mu^{-1}(\sqrt{\mu^2+1}-1)}\right) & \text{if } d=0, \\ 2(l(d;d)-l(\mu;d)) & \text{if } d\neq 0. \end{cases}$$

The null deviance for each observation is

$$D_0(d) = \begin{cases} 2\log\left(\frac{1+\mu^{-1}(\sqrt{\mu^2+1}-1)}{1-\mu^{-1}(\sqrt{\mu^2+1}-1)}\right) & \text{if } d = 0, \\ 2(l(d;d) - l(\hat{\mu};d)) & \text{if } d \neq 0, \end{cases}$$

where $\hat{\mu}$ is the mean of the $|d|$'s.

Parameter estimation

From the theory of exponential families (Azzalini, 1996), for a sample $\{d_i\}_{i=1}^n$ of independent and identically distributed variables following the probability mass function $f(d;p)$ as defined in Equation (5), the maximum likelihood estimator of $\mu = \mathbf{E}[|D|]$ is

$$\hat{\mu} = n^{-1} \sum_{i=1}^{n} |d_i|,$$

resulting in the maximum likelihood estimator of p

$$\hat{p} = \hat{\mu}^{-1}(\sqrt{\hat{\mu}^2 + 1} - 1)$$

by using Equation (6).

A generalized linear model

With these tools, we can easily define a generalised linear model. This is quite useful, e.g. in `R` (R Development Core Team, 2013), where we can create a family and use the functionality of the `glm` function and its cousins like the prediction function `predict`.

3. Estimation of Y-STR haplotype frequencies

In this section, we show how the discrete Laplace family introduced in Section 2.5 can be applied within the field of forensic genetics.

As introduced in Section 2, the normalised allele process $V_g(i) = X_g(i) - X_g(N)$ is the allele difference between any individual n and a fixed individual N. It was empirically validated that the discrete Laplace distribution is an approximation to the distribution of the normalised allele process.

Caliebe *et al.* (2010) uses $X_g(N)$, the allele of the N'th individual, as a reference in the normalised allele process. Note that any other person's allele can be used instead. We choose the reference as the median of all the alleles for one-locus haplotypes (for more loci, it is a bit more complicated and will be treated below). Thus, using the discrete Laplace distribution is merely a qualified guess as the results in (Caliebe *et al.*, 2010) will probably not hold when using the median instead of a fixed individual because the median is expected to have lower variance. Below, in Section 3.7, we investigate how qualified the guess actually is.

3.1. *Statistical model*

Let $\mathrm{DL}(p, m)$ be a discrete Laplace model with dispersion parameter $0 < p < 1$, where we now introduce a location parameter $m \in \mathbb{Z}$. The probability mass function is then

$$f(d; p, m) = \left(\frac{1-p}{1+p}\right) p^{|d-m|}.$$

Inference for a sample, $\{d_i\}_{i=1}^n$, can be made by noticing that the MLE's (maximum likelihood estimates) are

$$\begin{aligned} \hat{m} &= \text{median}\{d_i\}_{i=1}^n, \\ \hat{\mu} &= \frac{1}{n}\sum_{i=1}^n |d_i - \hat{m}| \quad \text{and} \\ \hat{p} &= \hat{\mu}^{-1}\left(\sqrt{\hat{\mu}^2+1} - 1\right), \end{aligned}$$

where the equation of $\hat{p}$ stems from Equation (6).

We will now introduce a model to perform inference in a mixture of multivariate, marginally independent, discrete Laplace distributions.

3.2. *Statistical model for multivariate mixtures*

Remember that we have r loci instead of just one (mutations across loci are assumed to happen independently). We assume that we have a mixture of c unobserved subpopulations centered at $y_j = (y_{j1}, y_{j2}, \ldots, y_{jr})$ for $j = 1, 2, \ldots, c$. We then assume that given a subpopulation, the signed allele distances to the subpopulation center follow independent discrete Laplace distributions.

As before, let $f(d; p)$ be the probability mass function of a $\mathrm{DL}(p, 0)$ distribution. We define an observation $X = (X_1, X_2, \ldots, X_r)$ to be a mixture of multivariate, marginally independent, discrete Laplace distributions when the probability of observing $X = x$ is

$$\sum_{j=1}^{c} \tau_j \prod_{k=1}^{r} f\left(|x_k - y_{jk}|; p_{jk}\right),$$

where τ_j is the priori probability for originating from the j'th subpopulation. Thus, the parameters of this mixture model are $\{y_j\}_{j=1}^c$, $\{\tau_j\}_{j=1}^c$ and $\{p_{jk}\}_{\substack{j\in\{1,2,\ldots,c\}\\ k\in\{1,2,\ldots,r\}}}$.

Let $\mathrm{MMDL}\left(c, r, \{y_j\}_{j=1}^c, \{\tau_j\}_{j=1}^c, \{p_{jk}\}_{\substack{j\in\{1,2,\ldots,c\}\\ k\in\{1,2,\ldots,r\}}}\right)$ denote such a mixture of multivariate, marginally independent, discrete Laplace distributions.

More theory on finite mixture distributions is given in (Titterington *et al.*, 1987).

3.3. *Likelihood*

In this section, the likelihood of the model is introduced. Let $x_i = (x_{i1}, x_{i2}, \ldots, x_{ir})$ for $i = 1, 2, \ldots, n$ denote the n observed haplotypes from a MMDL distribution. For

individual i and subpopulation j, let

$$d_{ijk} = |x_{ik} - y_{jk}|$$

be the distance at the k'th locus to the unknown location y_{jk}.

Let z_i denote the (unobserved) subpopulation from which the i'th haplotype originated such that $z_i = j$ when the i'th haplotype descents from the j'th subpopulation. Let

$$v_{ij} = \begin{cases} 1 & \text{if } z_i = j, \\ 0 & \text{otherwise,} \end{cases}$$

such that $v_{i+} = \sum_{j=1}^{c} v_{ij} = 1$.

Let $\tau_j = P(z_i = j)$ denote the a priori probability of originating from the j'th subpopulation yielding the constraint $\sum_j \tau_j = 1$. We will soon see that τ_j can be estimated by $\hat{\tau}_j = \hat{v}_{+j}/n = \sum_{i=1}^{n} \hat{v}_{ij}/n$, where $\hat{v}_{ij}$ is an estimate of $P(v_{ij} = 1 \mid x_i)$.

The full likelihood of individual i is given by

$$\begin{aligned} P(x_i, z_i) &= \prod_{j=1}^{c} \left(P(z_i = j)P(x_i \mid z_i = j)\right)^{v_{ij}} \\ &= \prod_{j=1}^{c} \left(P(z_i = j) \prod_{k=1}^{r} f(d_{ijk}; p_{jk})\right)^{v_{ij}} \\ &= \prod_{j=1}^{c} \tau_j^{v_{ij}} \prod_{k=1}^{r} f(d_{ijk}; p_{jk})^{v_{ij}}, \end{aligned}$$

where $f(d_{ijk}; p_{jk})$ is the probability mass function of the discrete Laplace distribution. Note, that p_{jk} in this case is assumed to depend on locus and subpopulation. We will assume that $\log p_{jk} = \theta_{jk} = \omega_j + \lambda_k$. This means that there is an additive effect of locus and an additive effect of subpopulation and that they do not depend on each other as there is no interaction term. This can be interpreted as ω_j representing the age of the j'th subpopulation and λ_k representing the mutation rate at the k'th locus.

Hence, the full likelihood of the n independent observations $\{x_i\}_{i=1}^{n}$ is

$$\begin{aligned} L_f &= L_f\left(\{p_{jk}\}_{j,k}, \{y_j\}_j, \{\tau_j\}_j, \{v_{ij}\}_{i,j}; \{x_i\}_i\right) \\ &= \prod_{i=1}^{n} P(x_i, z_i) \\ &= \prod_{i=1}^{n} \prod_{j=1}^{c} \tau_j^{v_{ij}} \prod_{k=1}^{r} f(d_{ijk}; p_{jk})^{v_{ij}} \\ &= \prod_{i=1}^{n} \prod_{j=1}^{c} \prod_{k=1}^{r} \left(\tau_j^{1/r} f(d_{ijk}; p_{jk})\right)^{v_{ij}}, \end{aligned}$$

where $d_{ijk} = |x_{ik} - y_{jk}|$ and $\log p_{jk} = \omega_j + \lambda_k$.

The marginal likelihood of the observed data is

$$L_m = L_m\left(\{p_{jk}\}_{j,k}, \{y_j\}_j, \{\tau_j\}_j; \{x_i\}_i\right) \tag{7}$$

$$
\begin{aligned}
&= \prod_{i=1}^{n} P(x_i) \\
&= \prod_{i=1}^{n} \sum_{j=1}^{c} P(x_i \mid z_i = j) P(z_i = j) \\
&= \prod_{i=1}^{n} \sum_{j=1}^{c} \tau_j \prod_{k=1}^{r} f(d_{ijk}; p_{jk}).
\end{aligned}
$$

It is a problem that the value of v_{ij} is not known. To deal with this problem, we consider the v_{ij}'s as unobserved variables and use the EM algorithm (Dempster *et al.*, 1977) to estimate the v_{ij}'s.

3.4. *Choose subpopulation centers*

The simplest way to determine the subpopulation centers, $\{y_j\}_{j=1}^{c}$, is to choose c subpopulation centers and keep these fixed. A more flexible approach is to first choose the initial subpopulation centers, and then allow for the subpopulation centers to be moved around later on if that makes the model better.

Due to the single step mutation model, clustering minimising the L^1 norm is an obvious choice for initial subpopulation centers as the same mutation rate is assumed for all alleles. This type of clustering is also sometimes referred to as k-medians (the method called k-means is minimising the L^2 norm). One of the possible methods doing this is the Partitioning Around Medoids (PAM) algorithm (Kaufman and Rousseeuw, 1990), which is supplied by the `R` (R Development Core Team, 2013) library `cluster` (Maechler *et al.*, 2005).

A disadvantage of PAM is that the number of subpopulations must be specified beforehand, but one can use BIC (Schwarz, 1978) (Bayesian Information Criteria) to select the best number of subpopulations.

When initial subpopulation centers are chosen, the parameters of the model are estimated using the EM algorithm (Dempster *et al.*, 1977) as described in Section 3.5.

When the EM algorithm has converged, one can try to move the subpopulation centers. Let $\hat{v}_{ij}$ denote the estimate of $P(v_{ij} = 1 \mid x_i)$ after the EM algorithm has converged. Because loci are independent in terms of the mutation process, the total likelihood consists of a product of likelihoods for each locus. This means that we can look at each locus at a time. Let $k \in \{1, 2, \ldots, r\}$ be the locus that should be considered.

The MLE of the subpopulation center location assuming all other information is known is then given by

$$
\hat{y}_{jk} = \operatorname*{argmin}_{y=\min_i\{x_{ik}\}}^{\max_i\{x_{ik}\}} \sum_{i=1}^{n} \sum_{j=1}^{c} \hat{v}_{ij} |x_{ik} - y|,
$$

as $g(y) = \sum_{i=1}^{n} \sum_{j=1}^{c} \hat{v}_{ij} |x_{ik} - y|$ is a convex, piecewise linear function that only needs to be evaluated in the ends of each line segment in order to find its minimum.

3.5. *EM algorithm*

Recall that

$$\mathbf{E}[v_{ij} \mid x_i] = P(z_i = j \mid x_i) \quad \text{and}$$
$$\tau_j = P(z_i = j)$$

and that p depends on locus and subpopulation with no interaction such that $\log p_{jk} = \theta_{jk} = \omega_j + \lambda_k$.

In the following equation, let

$$\mathbf{E}_v := \mathbf{E}_{\{v_{ij}\}_{i,j} \mid \{x_i\}_i, \{y_j\}_j, \{\tau_j\}_j, \{p_{jk}\}_{j,k}}$$

such that

$$\begin{aligned}
\mathbf{E}_v\left[\log L_f\right] &= \mathbf{E}_v\left[\log\left(\prod_{i=1}^{n}\prod_{j=1}^{c}\prod_{k=1}^{r}\left(\tau_j^{\frac{1}{r}} f(d_{ijk}; p_{jk})\right)^{v_{ij}}\right)\right] \\
&= \mathbf{E}_v\left[\sum_{i=1}^{n}\sum_{j=1}^{c} v_{ij} \sum_{k=1}^{r}\log\left(\tau_j^{\frac{1}{r}} f(d_{ijk}; p_{jk})\right)\right] \\
&= \sum_{i=1}^{n}\sum_{j=1}^{c}\mathbf{E}\left[v_{ij} \mid \{x_i\}_{i=1}^{n}\right] \times \\
&\quad \sum_{k=1}^{r}\log\left(\tau_j^{\frac{1}{r}} f(d_{ijk}; p_{jk})\right).
\end{aligned}$$

To obtain an estimate of v_{ij}, note that

$$\begin{aligned}
\mathbf{E}[v_{ij} \mid x_i] &= P(z_i = j \mid x_i) \\
&= \frac{P(z_i = j)P(x_i \mid z_i = j)}{\sum_{l=1}^{c} P(z_i = l)P(x_i \mid z_i = l))} \\
&= \frac{\tau_j \prod_k f(d_{ijk}; p_{jk})}{\sum_l \tau_l \prod_k f(d_{ilk}; p_{lk})},
\end{aligned}$$

which gives

$$\hat{v}_{ij} = \frac{\hat{\tau}_j \prod_k f(d_{ijk}; \hat{p}_{jk})}{\sum_l \hat{\tau}_l \prod_k f(d_{ilk}; \hat{p}_{lk})}$$

by using the estimates $\hat{\tau}_j$ and $\hat{p}_{jk}$ of τ_j and p_{jk}, respectively. For easier notation, let

$$\hat{w}_{ij} = \hat{\tau}_j \prod_k f(d_{ijk}; \hat{p}_{jk}) \quad \text{and}$$

$$\hat{v}_{ij} = \frac{\hat{w}_{ij}}{\sum_l \hat{w}_{il}}. \tag{8}$$

And similar to earlier

$$\hat{\tau}_j = \frac{\hat{v}_{+j}}{n}, \tag{9}$$

where $\hat{v}_{+j} = \sum_{i=1}^{n} \hat{v}_{ij}$.

Now, the EM algorithm used can be described:

- **E-step**: Calculate $\hat{v}_{ij}$ using Equation (8) using the current estimates of $\hat{\tau}_j$ and $\hat{p}_{jk}$ (obtained from the previous E-step and M-step). Now, $\hat{\tau}_j$ can be updated using Equation (9).
- **M-step**: Maximise

$$L_f = \prod_{i=1}^{n} \prod_{j=1}^{c} \prod_{k=1}^{r} \left(\tau_j^{1/r} f(d_{ijk}; p_{jk}) \right)^{v_{ij}}$$

for $\{p_{jk}\}_{j,k}$ using the current estimates for the other parameters:

$$\begin{aligned} \{\hat{p}_{jk}\}_{j,k} &= \underset{\{p_{jk}\}_{j,k}}{\arg\max} L_f \\ &= \underset{\{p_{jk}\}_{j,k}}{\arg\max} \prod_{i=1}^{n} \prod_{j=1}^{c} \prod_{k=1}^{r} \left(\hat{\tau}_j^{1/r} f(d_{ijk}; p_{jk}) \right)^{\hat{v}_{ij}} \\ &= \underset{\{p_{jk}\}_{j,k}}{\arg\max} \prod_{i=1}^{n} \prod_{j=1}^{c} \prod_{k=1}^{r} \left(f(d_{ijk}; p_{jk}) \right)^{\hat{v}_{ij}}. \end{aligned}$$

This can be done by assuming the GLM model $d_{ijk} \sim \omega_j + \lambda_k$ (other possibilities do exist) with weights $\hat{v}_{ij}$, where $p_{jk} = \exp(\omega_j + \lambda_k)$ (ω_j is a subpopulation effect corresponding to age and λ_k a locus effect corresponding to mutation rate), thus obtaining $\hat{p}_{jk}$.

The assumption that the power $\hat{v}_{ij}$ is equivalent to fixed, known weights in a GLM likelihood is shown in more detail in (Wedel and DeSarbo, 1995). The `R` (R Development Core Team, 2013) package `FlexMix` (Leisch, 2004; Grün and Leisch, 2008) also uses the same strategy to fit mixtures of GLMs.

According to (Dempster *et al.*, 1977, Theorem 1, p. 7), the marginal likelihood Equation (7) increases with each step of the EM algorithm. Starting values can be chosen as

$$\hat{\tau}_j = 1/c \quad \text{and} \quad \hat{\mu}_{ijk} = d_{ijk} + 0.1,$$

where $\hat{\mu}_{ijk}$ is chosen such that the boundary is avoided.

This EM algorithm making inference in a MMDL distribution (mixture of multivariate, marginally independent, discrete Laplace distributions) was implemented in the `R` (R Development Core Team, 2013) package `disclapmix` (Andersen and Eriksen, 2013b).

Note, that there are $cr + (r + c - 1) + (c - 1)$ parameters in a MMDL distribution: cr for the subpopulation centers,

$$\{y_j\}_{j=1}^{c};$$

$(r+c-1)$ for the parameters in the multivariate, marginally independent, discrete Laplace distributions,

$$\{p_{jk}\}_{\substack{j \in \{1,2,\ldots,c\} \\ k \in \{1,2,\ldots,r\}}},$$

as there are only main effects of subpopulation and locus; and $c-1$ for the prior probabilities,

$$\{\tau_j\}_{j=1}^{c},$$

of originating from each of the c subpopulations, with the reduction of 1 parameter as they sum to 1.

3.6. *Haplotype frequency prediction*

Given subpopulation centers $\{\hat{y}_j\}_j$, parameters $\{\hat{p}_{jk}\}_{j,k}$ and prior probabilities $\{\hat{\tau}_j\}_j$, e.g. from a converged run of the EM algorithm described in Section 3.5, the haplotype frequency of a haplotype $h = (h_1, h_2, \ldots, h_r)$ with $h_k \in \mathbb{Z}$ for $k \in \{1, 2, \ldots, r\}$ can be estimated as

$$\sum_{j=1}^{c} \hat{\tau}_j \prod_{k=1}^{r} f\left(|h_k - \hat{y}_{jk}|; \hat{p}_{jk}\right).$$

3.7. *Simulation study*

To assess the model described in Section 3 for estimating Y-STR (a haploid lineage DNA marker) haplotype frequencies, a simulation study was performed.

A population under the Fisher-Wright model (Fisher, 1922, 1930, 1958; Wright, 1931; Ewens, 2004) with a neutral (in terms of no selection), single step mutation process (Ohta and Kimura, 1973) was simulated using the `R` (R Development Core Team, 2013) package `fwsim` (submitted, see Andersen and Eriksen (2012a) for a preprint). The datasets from this population were sampled and used for estimating haplotype frequencies that were compared to the population frequency.

We simulated 12 different population types by taking all possible combinations of

- Loci: $r = 7$
- Mutation rate: $\mu = 0.01$, 0.003 or 0.001
- Generations: $g = 500$ or 1,000
- Initial population size: $k = 10{,}000$ or 50,000.

For all types, the resulting expected population size after g generations was 20,000,000 due to a constant population growth, ρ, that was determined using the number of generations and initial population size as follows. Let N_i denote the population size at the i'th generation. The model from `fwsim` assumes that $N_{i+1} \mid N_i \sim \text{Poisson}(\rho N_i)$. Thus, if g denotes the number of generations (500 or 1,000) and N_0 the initial population size (10,000 or 50,000), then $\mathbf{E}[N_g] = \rho^g N_0$.

For each combination of the parameters, five realisations of the population were simulated. For each of these populations, 50 datasets of size 500, 1,000 and

5,000 were drawn. In total, $12 \cdot 5 \cdot 3 \cdot 50 = 9{,}000$ datasets were sampled and used as basis for comparison.

Note, that the simulated populations are idealised in the sense that the match probability is the haplotype frequency. For all singletons in the dataset, the discrete Laplace distribution approach described in Section 3 was compared to the naïve $1/(n+1)$ estimator and to Brenner's $(1-\kappa)/(n+1)$ estimate, where $\kappa = (\alpha+1)/(n+1)$ and $\alpha+1$ is the number of singletons (haplotypes observed only once) in the dataset as inspired by Robbins (1968). As previously mentioned, the discrete Laplace distribution approach described in Section 3 is implemented in the `R` package `disclapmix` (Andersen and Eriksen, 2013b) that can be used as follows:

```
library(disclapmix)

# Load the dataset danes
data(danes)

# The dataset consists of the haplotype and
# the number of times it has been observed
head(danes)

# Make a dataset consisting of one observation per row
db <- as.matrix(danes[rep(1:nrow(danes), danes$n), 1:(ncol(danes)-1)])

# Fit the model with up to 5 subpopulations
clusters <- 1L:5L # L to force integer
res <- lapply(clusters, function(clusters)
  disclapmix(db, clusters = clusters))

# See the most important information
marginalBICs <- sapply(res, function(fit) fit$BIC_marginal)
bestfit <- res[[which.min(marginalBICs)]]

# Predict haplotype frequencies
disclap.estimates <- predict(bestfit, newdata = as.matrix(danes[, 1:10]))
```

For futher information on functionality and usage, please run `demo(danes)` and refer to the documentation `?disclapmix`.

As performance measures, the observed bias and the Kullback-Leibler divergence (Kullback and Leibler, 1951; Kullback, 1959) were calculated. Because it is most problematic to estimate the frequency of singletons (haplotypes only observed once), we only focus on these. For a haplotype dataset $H = \{h_i\}_{i=1}^n$ with singletons $\{h_i\}_{i\in S}$ and population frequencies $\{p_i\}_{i\in S}$ estimated as $\{P_{E(H)}(h_i)\}_{i\in S}$ by an estimator E, the bias is

$$B_{H,S}(E) = \frac{1}{|S|}\sum_{i\in S}(P_{E(H)}(h_i) - p_i). \tag{10}$$

The Kullback-Leibler divergence is a measure in information theory about the distance between two probability distributions (we used this distance in Section 2.4) and can also be interpreted as a prediction error. In this case, we only have binary probability distributions. If a haplotype has population frequency p

and is estimated to $\hat{p}$, then the Kullback-Leibler divergence is

$$D_{KL}(\hat{p}; p) = \hat{p} \log\left(\frac{\hat{p}}{p}\right) + (1 - \hat{p}) \log\left(\frac{1 - \hat{p}}{1 - p}\right).$$

The distribution of Kullback-Leibler divergences for singletons $\{h_i\}_{i \in S}$ is

$$D_{H,S}(E) = \{D_{KL}(P_{E(H)}(h_i); p_i)\}_{i \in S}. \tag{11}$$

The mean and upper 95% quantile of the distribution of Kullback-Leibler divergences for the naïve $1/(n+1)$ estimator, Brenner's κ estimator, and the discrete Laplace based estimator were compared together with the bias.

Note, that the lowest possible prediction error in terms of the Kullback-Leibler divergence is 0, which occurs when $\hat{p} = p$. If this happens for all singletons – that is, all singletons' frequencies were perfectly estimated – then the mean of the Kullback-Leibler divergences would be 0 and so would the bias be. Hence, if the mean of Kullback-Leibler divergence is 0, then so is the bias.

On the other hand, if the bias is 0, then we do not know anything about the Kullback-Leibler divergences. The bias could be 0 if either all the singletons' frequencies were perfectly estimated or if some frequencies were somehow overestimated and others were equally underestimated such that they cancelled each other out.

Thus, the prediction error is telling us about the size of the error, whereas the bias is telling us about the direction of the error.

Because migration was not included in the simulation of the populations, only one subpopulation for the discrete Laplace based estimator was used.

Results

As naming convension, DiscLap refers to the model described in Section 3.

For all population types in our simulation study and the performance measures mentioned, the naïve $1/(n+1)$ estimator performed much worse than Brenner's κ estimator and the DiscLap estimator.

Figure 5 shows estimation in a single dataset (one out of the 9,000 datasets analysed in total). Figure 6 shows the singleton proportions for the simulated datasets.

The bias as defined in Equation (10) is shown in Figure 7. Both the naïve estimator and Brenner's κ estimator seem, in general, to be conservative, which is also what Brenner (2010) states. For dataset size 500, DiscLap seems almost unbiased.

This tendency seems stronger for dataset sizes of 1,000 and 5,000. For the low mutation rate of 0.001, DiscLap seems slightly anti-conservative, whereas for the higher mutation rate of 0.003, it almost seems to be unbiased.

When it comes to the distribution of Kullback-Leibler divergences as defined in Equation (11), Figure 8 (the mean) and Figure 9 (the upper 95% quantile) show the same picture, namely that DiscLap overall seems better than Brenner's κ estimator. Table 1 shows a summary of the average proportion between Brenner's κ and DiscLap of the mean of the Kullback-Leibler divergences for each mutation rate and database size.

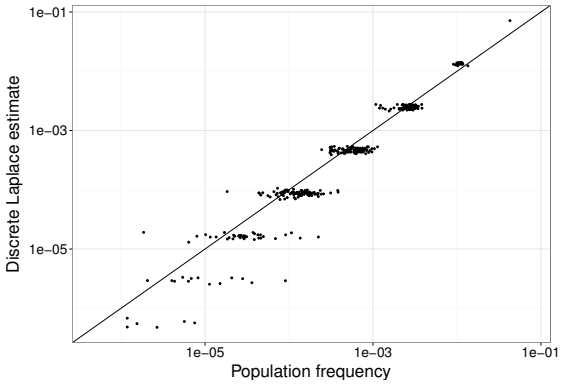


Figure 5. Haplotype singleton frequency estimation in a single dataset of size 500 from a population with an initial size of 10,000 evolved in 500 generations, a mutation rate of 0.001 and a population growth leading to an expected population size of 20,000,000 after 500 generations. The actual end population size was 19,397,385 consisting of 34,180 different haplotypes.

	$\mu = 0.001$	$\mu = 0.003$	$\mu = 0.01$
$n = 500$	23.60	4.88	34.22
$n = 1{,}000$	18.67	3.72	5.71
$n = 5{,}000$	9.54	4.01	0.86

Table 1. The average proportion between Brenner's κ and DiscLap of the mean of the Kullback-Leibler divergences for database summarised by mutation rate μ and database size n. A proportion greater than 1 means that the mean of the Kullback-Leibler divergences for Brenner's κ was higher than that of DiscLap. And opposite for a proportion lower than 1.

Discussion

In summary, the prediction error of the estimator using the discrete Laplace distribution (DiscLap) was lower than those of both the κ model by Brenner (2010) and the naïve $1/(n+1)$ estimator. For all population types in our simulation study and the performance measures mentioned (bias and Kullback-Leibler divergence), the naïve $1/(n+1)$ estimator performed much worse than Brenner's κ estimator and the DiscLap estimator.

It seems as if Brenner's κ model estimates haplotype frequencies rather well although it does not incorporate genetic information. One major drawback of this method is that all unobserved haplotypes are assigned the same frequency estimate. Hence, it is doubtful if Brenner's κ model for example is suitable to separate a mixture based on calculating the likelihood ratio (LR) as a measure of the weight of evidence.

Another really important difference between Brenner's κ model and DiscLap is that DiscLap is also able to estimate frequencies for non-singleton haplotypes. Thus, DiscLap can be used no matter if the haplotype has been observed before or not.

In the population types that we studied, we did not observe situations where

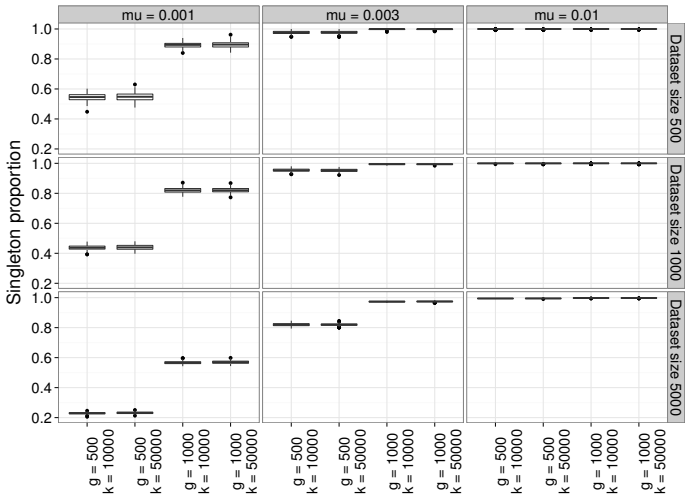


Figure 6. Singleton proportions of the 9,000 simulated datasets.

the estimator based on the discrete Laplace distribution performed worse than the estimator based on Brenner's κ model.

We encourage research on how different population models and migration affects Brenner's κ model and the discrete Laplace distribution.

3.8. *Real data example*

We analysed the 1,774 German 17-marker haplotypes from release 37 of the YHRD `http://www.yhrd.org` (Roewer *et al.*, 2001; Willuweit and Roewer, 2009). To render the data usable for both discrete Laplace estimation and the frequency surveying method (Roewer *et al.*, 2000; Krawczak, 2001; Willuweit *et al.*, 2011), some markers and haplotypes were excluded. First, `DYS385a/b` was ignored because of its inherent genotype ambiguity (Roewer *et al.*, 2000) leaving 15 markers for further analysis. Next, four haplotypes with two alleles reported at `DYS19` and 13 haplotypes with incomplete repeats were excluded, leaving n = 1,757 haplotypes in the data set. Finally, alleles at `DYS389II` were replaced by `DYS389II` minus `DYS389I` (Butler, 2005). Out of the 1,757 haplotypes analysed, 1,469 were singletons.

When restricting the genotype information to the so-called 'minimal haplotype' comprising the seven loci `DYS19`, `DYS389I`, `DYS389II`, `DYS390`, `DYS391`, `DYS392`, and `DYS393`, a total of 392 singletons were observed among the haplotypes of the German data.

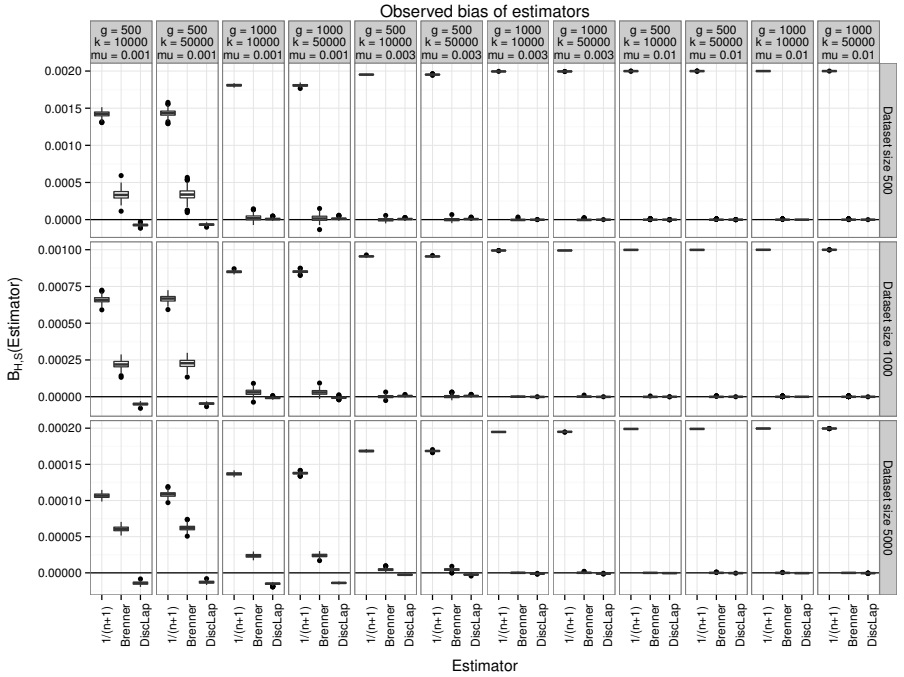


Figure 7. Bias for the different estimators as defined in Equation (10).

Frequency surveying

In its revised form, the surveying method (Willuweit *et al.*, 2011) was based upon an exponential regression model

$$\mu_i = \exp(r_1 W_i + r_2) \quad \text{and} \quad \sigma_i = \exp(s_1 W_i + s_2)$$

that links the mean, μ_i, and the standard deviation, σ_i, of the population frequency of the i'th haplotype to its weighted inverse molecular distance, W_i, from all other haplotypes in the database. Once the regression parameters, r_1, r_2, s_1, s_2, were determined, the model could serve to define a prior beta distribution of the frequency of any haplotype, h_0, with molecular distance W_0. The parameters of this prior distribution were calculated as

$$\alpha_0 = \frac{\mu_0^2(1-\mu_0)}{\sigma_0^2} - \mu_0 \quad \text{and} \quad \beta_0 = \alpha_0\left(\frac{1-\mu_0}{\mu_0}\right).$$

Maximum likelihood estimates of the regression parameters were obtained in our study by numerical optimisation (Willuweit *et al.*, 2011) using the Nelder-Mead simplex algorithm with up to 1,500 iterations as implemented in `R` (R Development Core Team, 2013). Several different starting values of (r_1, r_2, s_1, s_2) were tried, and the vector resulting in the highest likelihood

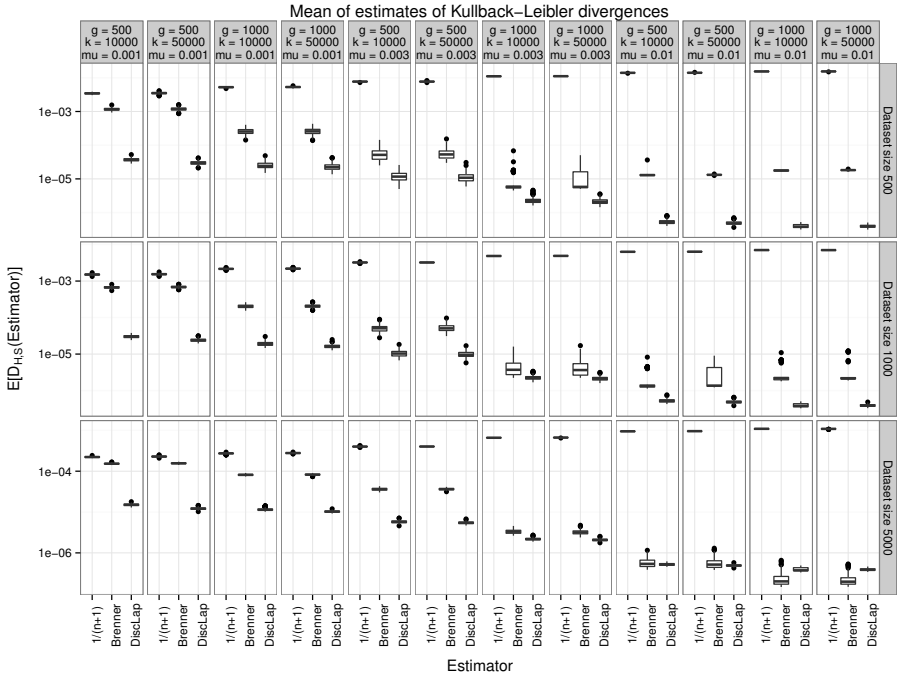


Figure 8. Mean of the Kullback-Leibler divergences defined in Equation (11) for each population type. Note, that the ordinate is on a log scale.

was chosen. The starting values were taken from the Cartesian product $\{15,20,30.82\} \times \{-10,-15,-13.17\} \times \{15,20,28.95\} \times \{-10,-15,-11.71\}$, where the last elements in the sets are the respective binning estimates of the Western European population given in Table 3 of (Willuweit *et al.*, 2011).

Let n_i be the number of times that the i'th haplotype was observed in the database with $n = \sum_i n_i$ being equal to the database size. For comparison with the other estimators, we used the mean of the posterior $\text{Beta}(\alpha_i + n_i - 1, \beta_i + n - n_i)$ given by

$$\frac{\alpha_i + n_i - 1}{\alpha_i - 1 + \beta_i + n}$$

as the haplotype surveying estimate of the population frequency of haplotype, h_i.

Results

For both the full and the minimal haplotype, only the singletons were used to compare the haplotype frequency estimates provided by the different estimators.

Figure 10 shows the results of the 7-loci-database. Figure 11 shows the results of the 15-loci-database. It is impossible to make any sensible conclusion from this as we do not know the true haplotype frequencies.

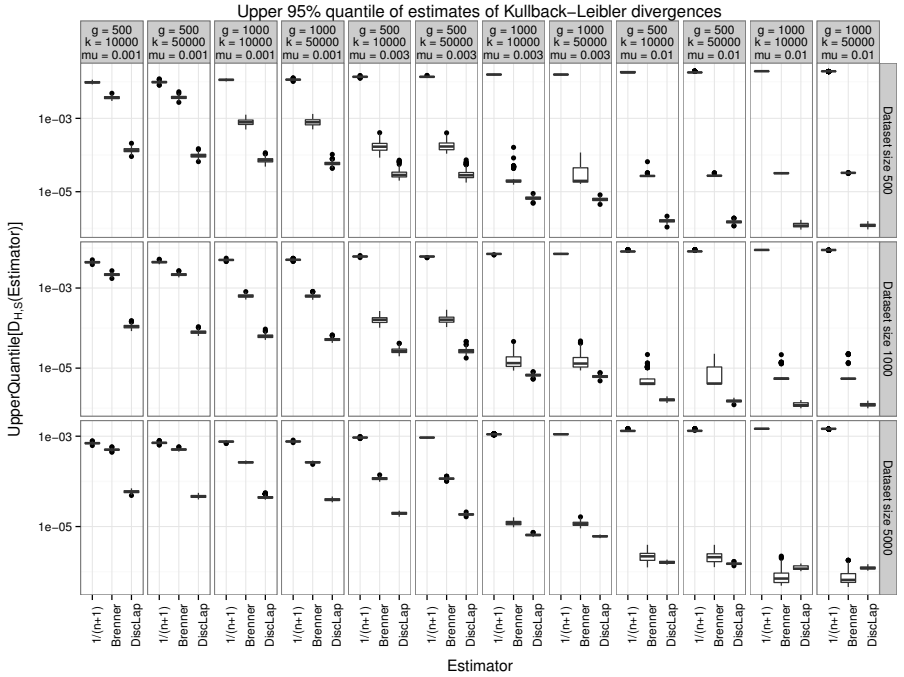


Figure 9. Upper 95% quantile of the Kullback-Leibler divergences defined in Equation (11) for each population type. Note, that the ordinate is on a log scale.

4. Discussion

The first part of this paper describes an exponential family called the discrete Laplace distribution. The fact that the discrete Laplace distribution is an exponential family makes inference somewhat easier as theory on exponential families already exists and can be exploited. This also means simpler and faster computer software because existing implementations that have been optimised can be used.

The second part of this paper consists of an application of the discrete Laplace distribution, namely how to estimate Y-STR haplotype frequencies. An estimate of the frequency of a Y-STR haplotype can be used as an estimate of the match probability (assuming an idealised population without population substructure), which is an essential part in forensic genetics when evaluating the evidential weight of the evidence by means of likelihood principles. The calculations could be performed on a normal computer. We demonstrate that for our simulation study on 12 different population types (varying mutation rate, population growth and generations) resulting in 9,000 datasets (of size 500, 1,000 and 1,500), the haplotype frequency estimation based on the discrete Laplace distribution performs overall better than the κ model by Brenner (2010). The mean of the Kullback-Leibler divergences is in general lower for the estimation based on the

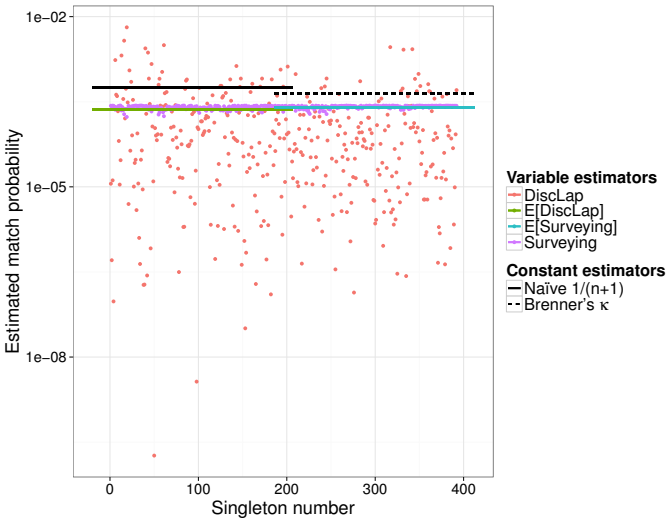


Figure 10. Comparison of the haplotype frequency estimators for the 7 loci German database consisting of 1,757 haplotypes of which 392 were singletons. Thus, Brenner's $\kappa = 4.4 \cdot 10^{-4}$. Note, that the ordinate with the estimated haplotype frequency is on a log scale. 11 subpopulations were used (1 through 15 subpopulations were tried, 11 subpopulations had the lowest BIC score (Schwarz, 1978)). The line 'E[DiscLap]' refers to the average of the DiscLap estimates and the line 'E[Surveying]' refers to the average of the surveying estimates.

discrete Laplace distribution than that based on Brenner's κ cf. Table 1.

Furthermore and very importantly, Brenner's κ can only be used for singletons whereas estimation based on the discrete Laplace distribution can be used for all haplotypes.

We encourage research on how different population models and migration affects Brenner's κ model and the discrete Laplace distribution.

5. Acknowledgements

We thank Amke Caliebe, Christian-Albrechts-Universität zu Kiel, Germany, for providing the `R` code needed to estimate haplotype frequencies using the surveying approach.

6. Bibliography

Andersen, M. M. (2010) *Y-STR: Haplotype Frequency Estimation and Evidence Calculation*. Master's thesis, Aalborg University, Denmark. 78

Andersen, M. M., Caliebe, A., Jochens, A., Willuweit, S. and Krawczak, M. (2013a) Estimating trace-suspect match probabilities for singleton Y-STR haplotypes using coalescent theory. *Forensic Science International: Genetics*, **7**, 264–271. 78, 80

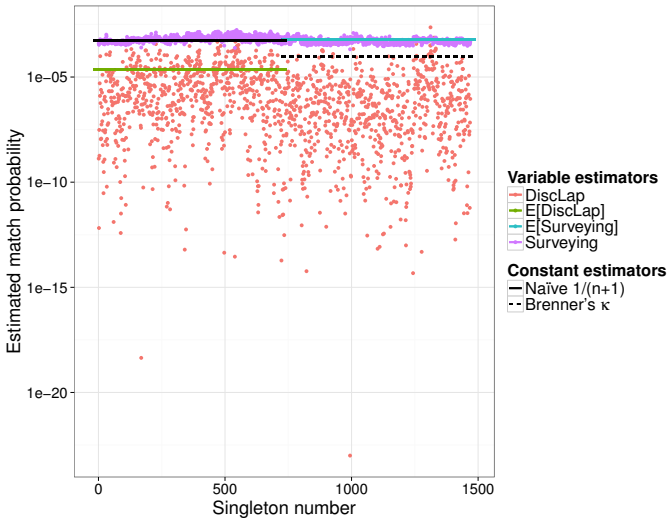


Figure 11. Comparison of the haplotype frequency estimators for the 15 loci German database consisting of 1,757 haplotypes of which 1,469 were singletons. Thus, Brenner's $\kappa = 9.3 \cdot 10^{-5}$. Note, that the ordinate with the estimated haplotype frequency is on a log scale. 14 subpopulations were used (1 through 15 subpopulations were tried, 14 subpopulations had the lowest BIC score (Schwarz, 1978)). The line 'E[DiscLap]' refers to the average of the DiscLap estimates and the line 'E[Surveying]' refers to the average of the surveying estimates.

Andersen, M. M. and Eriksen, P. S. (2012a) Efficient Forward Simulation of Fisher-Wright Populations with Stochastic Population Size and Neutral Single Step Mutations in Haplotypes. *Preprint, arXiv:1210.1773*. 79, 93

Andersen, M. M. and Eriksen, P. S. (2012b) *fwsim: Fisher-Wright Population Simulation*. URL `http://CRAN.R-project.org/package=fwsim`. R package version 0.2-5. 79

Andersen, M. M. and Eriksen, P. S. (2013a) *disclap: Discrete Laplace Family*. URL `http://CRAN.R-project.org/package=disclap`. R package version 1.4. 79, 83

Andersen, M. M. and Eriksen, P. S. (2013b) *disclapmix: Discrete Laplace mixture inference using the EM algorithm*. URL `http://CRAN.R-project.org/package=disclapmix`. R package version 1.2. 79, 92, 94

Andersen, M. M., Eriksen, P. S. and Morling, N. (2013b) A gentle introduction to the discrete Laplace method for estimating Y-STR haplotype frequencies. *Preprint, arXiv:1304.2129*. 79

Azzalini, A. (1996) *Statistical Inference – Based on the Likelihood*. Chapman & Hall. 87

Ballantyne, K. N. *et al.* (2010) Mutability of Y-Chromosomal Microsatellites: Rates, Characteristics, Molecular Bases, and Forensic Implications. *The American Journal of Human Genetics*, **87**, 341–353. 82

Brenner, C. H. (2010) Fundamental problem of forensic mathematics – The evidential value of a rare haplotype. *Forensic Science International: Genetics*, **4**, 281–291. 78, 95, 96, 100

Brigham, E. O. (1988) *The fast Fourier transform and its applications*. Prentice Hall. 81

Buckleton, J., Krawczak, M. and Weir, B. (2011) The interpretation of lineage markers in forensic DNA testing. *Forensic Science International: Genetics*, **5**, 78–83. 78

Budowle, B., Aranda, X. *et al.* (2008) Null allele sequence structure at the DYS448 locus and implications for profile interpretation. *International Journal of Legal Medicine*, **122**, 421–427. 80

Butler, J. M. (2005) *Forensic DNA Typing: Biology, Technology, and Genetics of STR Markers*. Academic Press, 2nd edn. 80, 97

Caliebe, A., Jochens, A., Krawczak, M. and Rösler, U. (2010) A Markov Chain Description of the Stepwise Mutation Model: Local and Global Behaviour of the Allele Process. *Journal of Theoretical Biology*, **266**, 336–342. 79, 80, 81, 82, 87

Cooley, J., Lewis, P. and Welch, P. (1969) The finite Fourier transform. *IEEE Trans. Audio Electroacoustics*, **17**, 77–85. 81

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, **39**, 1–38. 79, 90, 92

Evett, I. W. and Weir, B. S. (1998) *Interpreting DNA Evidence*. Sinauer Associates. 78

Ewens, W. J. (2004) *Mathematical Population Genetics*. Springer-Verlag. 79, 93

Fisher, R. A. (1922) On the Dominance Ratio. *Proc. Roy. Soc. Edin.*, **42**, 321–341. 79, 93

Fisher, R. A. (1930) *The Genetical Theory of Natural Selection*. Oxford: Clarendon Press. 79, 93

Fisher, R. A. (1958) *The Genetical Theory of Natural Selection*. New York: Dover, 2nd revised edn. 79, 93

Gill, P., Brenner, C. *et al.* (2001) DNA Commission of the International Society of Forensic Genetics: Recommendations on forensic analysis using Y-chromosome STRs. *Forensic Science International*, **124**, 5–10. 78

Gill, P., Jeffreys, A. J. and Werrett, D. J. (1985) Forensic application of DNA fingerprints. *Nature*, **318**, 577–579. 78

Grün, B. and Leisch, F. (2008) FlexMix Version 2: Finite Mixtures with Concomitant Variables and Varying and Constant Parameters. *Journal of Statistical Software*, **28**. 92

Hein, J., Schierup, M. H. and Wiuf, C. (2005) *Gene Genealogies, Variation and*

Evolution: A Primer in Coalescent Theory. Oxford University Press. 80

Inusah, S. and Kozubowski, T. J. (2006) A discrete analogue of the Laplace distribution. *Journal of Statistical Planning and Inference*, **136**, 1090–1102. 79

Kaufman, L. and Rousseeuw, P. J. (1990) *Finding groups in data: An introduction to cluster analysis*. Wiley. 90

Krawczak, M. (2001) Forensic evaluation of Y-STR haplotype matches: a comment. *Forensic Science International*, **118**, 114–115. 78, 97

Kullback, S. (1959) *Information theory and statistics*. John Wiley and Sons. 82, 94

Kullback, S. and Leibler, R. A. (1951) On Information and Sufficiency. *Annals of Mathematical Statistics*, **22**, 79–86. 82, 94

Leisch, F. (2004) FlexMix: A General Framework for Finite Mixture Models and Latent Class Regression in R. *Journal of Statistical Software*, **11**. 92

Maechler, M., Rousseeuw, P., Struyf, A. and Hubert, M. (2005) Cluster analysis basics and extensions. 90

Ohta, T. and Kimura, M. (1973) A Model of Mutation Appropriate to Estimate the Number of Electrophoretically Detectable Alleles in a Finite Population. *Genet. Res.*, **22**, 201–204. 79, 93

R Development Core Team (2013) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org`. ISBN 3-900051-07-0. 79, 82, 83, 84, 87, 90, 92, 93, 98

Robbins, H. E. (1968) Estimating the Total Probability of the Unobserved Outcomes of an Experiment. *The Annals of Mathematical Statistics*, **39**, 256–257. 78, 94

Roewer, L. (2009) Y chromosome STR typing in crime casework. *Forensic Sci Med Pathol*, **5**, 77–84. 78

Roewer, L., Kayser, M., de Knijff, P. *et al.* (2000) A new method for the evaluation of matches in non-recombining genomes: application to Y-chromosomal short tandem repeat (STR) haplotypes in European males. *Forensic Science International*, **114**, 31–43. 78, 97

Roewer, L., Krawczak, M., Willuweit, S. *et al.* (2001) Online reference database of European Y-chromosomal short tandem repeat (STR) haplotypes. *Forensic Science International*, **2-3**, 106–113. 97

Schwarz, G. (1978) Estimating the Dimension of a Model. *Annals of Statistics*, **6**, 461–464. 90, 101, 102

Sibille, I., Duverneuil, C. *et al.* (2002) Y-STR DNA amplification as biological evidence in sexually assaulted female victims with no cytological detection of spermatozoa. *Forensic Sci. Int.*, **125**, 212–216. 78

Titterington, D. M., Smith, A. F. M. and Makov, U. E. (1987) *Statistical Analysis of Finite Mixture Distributions*. Wiley. 88

Wedel, M. and DeSarbo, W. S. (1995) A Mixture Likelihood Approach for Generalized Linear Models. *Journal of Classification*, **12**, 21–55. 92

Willuweit, S., Caliebe, A., Andersen, M. M. and Roewer, L. (2011) Y-STR Frequency Surveying Method: A critical reappraisal. *Forensic Science International: Genetics*, **5**, 84–90. 78, 97, 98, 99

Willuweit, S. and Roewer, L. (2009) Y chromosome haplotype reference database (YHRD): Update. *Forensic Science International: Genetics*, **1**, 83–87. 97

Wright, S. (1931) Evolution in Mendelian populations. *Genetics*, **16**, 97–159. 79, 93

Paper VII

A gentle introduction to the discrete Laplace method for estimating Y-STR haplotype frequencies

Author list Mikkel Meyer Andersen, *Aalborg University, Denmark*
Poul Svante Eriksen, *Aalborg University, Denmark*
Niels Morling, *University of Copenhagen, Denmark*

Summary Y-STR data simulated under a Fisher-Wright model of evolution with a single-step mutation model turns out to be well predicted by a method using discrete Laplace distributions.

Publication info This paper has been made publicly available as:

Andersen MM, Eriksen PS, Morling N (2013). *A gentle introduction to the discrete Laplace method for estimating Y-STR haplotype frequencies.* arXiv: 1304.2129.

1. Introduction

This tutorial introduces the discrete Laplace method for estimating Y-STR haplotype frequencies as described by Andersen *et al.* (2013).

To accomplish this, we demonstrate a number of examples using `R` (R Development Core Team, 2013). The code examples look like the following that loads the `disclap` package (Andersen and Eriksen, 2013a) which is needed for the following examples:

```
> library(disclap)
```

If you do not have installed the `disclap` package, please visit `http://cran.r-project.org/package=disclap`.

2. The discrete Laplace distribution

The discrete Laplace distribution is a probability distribution like e.g. the binomial distribution or the normal/Gaussian distribution.

The discrete Laplace distribution has two parameters: a dispersion parameter $0 < p < 1$ and a location parameter $y \in \mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$.

Let $X \sim DL(p, y)$ denote that the random variable X follows a discrete Laplace distribution with dispersion parameter $0 < p < 1$ and location parameter y. Then a realisation of the random variable, $X = x$, can be any integer in $\mathbb{Z}$. The random variable X has the probability mass function given by

$$f(X = x; p, y) = \frac{1-p}{1+p} \cdot p^{|x-y|} \quad \text{for } x \in \mathbb{Z}.$$

As seen, only the absolute value of $x - y$ is used. This means that the probability mass function is symmetric around y.

Let us try to plot the probability mass function $f(X = x; p, y)$ for $p = 0.3$ and $y = 13$ from $x = 8$ to $x = 18$:

```
> p <- 0.3
> y <- 13
> x <- seq(8, 18, by = 1)
> barplot(ddisclap(x - y, p), names = x, xlab = "x, e.g. Y-STR allele",
    ylab = paste("Probability mass, f(X = x; ", p, ", ", y, ")", sep = ""))
```

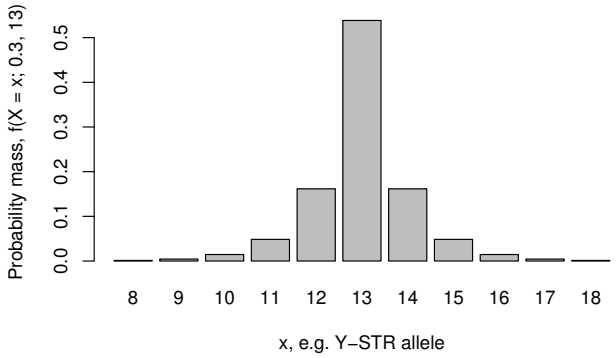


Figure 1. The probability mass function, $f(X = x; p, y)$, for the discrete Laplace distribution with dispersion parameter $p = 0.3$ and location parameter $y = 13$ from $x = 8$ to $x = 18$.

We plot the distribution for values of x from 8 to 18 as there is almost no probability mass outside these values. We can find out how much of the probability mass that we have plotted:

```
> sum(ddisclap(abs(x - y), p))
[1] 0.9989
```

Thus, only 0.0011 of the probability mass is outside $\{8, 9, \ldots, 17, 18\}$.

If we have a sample of realisations from $X \sim DL(p, y)$ denoted by $\{x_i\}_{i=1}^n$, then maximum likelihood estimates are given by the following quantities (Andersen *et al.*, 2013):

$$\hat{y} = \text{median}\{x_i\}_{i=1}^n,$$
$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} |x_i - \hat{y}| \text{ and}$$
$$\hat{p} = \hat{\mu}^{-1}\left(\sqrt{\hat{\mu}^2 + 1} - 1\right).$$

Example:

```
> set.seed(1)                 # Makes it possible to reproduce the results
> p <- 0.3                    # Dispersion parameter
> y <- 13                     # Location parameter
> x <- rdisclap(100, p) + y  # Generate a sample using the rdisclap
> y.hat <- median(x)
> y.hat
[1] 13
> mu.hat <- mean(abs(x - y.hat))
> mu.hat
[1] 0.57
> p.hat <- mu.hat^(-1) * (sqrt(mu.hat^2 + 1) - 1)
> p.hat  # We expect 0.3
```

```
[1] 0.265
> # The observed distribution of d's
> tab <- prop.table(table(x))
> tab
x
  10   11   12   13   14   15   16
0.01 0.03 0.15 0.55 0.20 0.05 0.01
```

This can be plotted against the expected counts as follows:

```
> plot(1:length(tab), ddisclap(as.integer(names(tab)) - y.hat, p.hat),
>   type = "h", col = "#999999", lend = "butt", lwd = 50,
>   xlab = "x, e.g. Y-STR allele", ylab = "Probability mass", axes = FALSE)
> axis(1, at = 1:length(tab), labels = names(tab))
> axis(2)
> points(1:length(tab), tab, type = "h", col = "#000000",
>   lend = "butt", lwd = 25)
> legend("topright", c("Estimated distribution", "Observations"),
>   pch = 15, col = c("#999999", "#000000"))
```

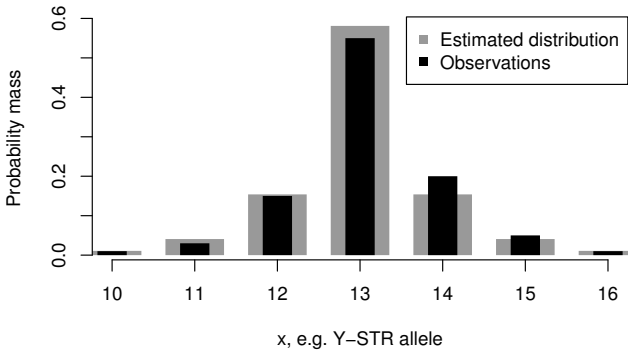


Figure 2. Observed frequencies of the x's compared to a discrete Laplace distribution with parameters estimated from the sample.

3. Mixtures of multivariate discrete Laplace distributions

Assume a very simple 'haplotype' with only one locus. Also assume a simple and isolated population. Then, it is reasonable to assume that there is a modal/central Y-STR allele, y, and that all the alleles are distributed around this allele.

If we go back to Figure 2, this can be illustrated by $y = 13$ as the central Y-STR allele and a distribution around $y = 13$ with shorter and longer alleles.

To begin with, it might seem a bit overwhelming that Y-STR alleles should follow a simple probabiity distribution such as the discrete Laplace distribution.

But surprisingly, it is actually a good approximation as demonstrated by Andersen *et al.* (2013).

We have haplotypes with several loci. When we assess multiple loci haplotypes, we assume that mutations happen independently across loci. Each locus has its own discrete Laplace distribution of allele probabilities, and the probability of a haplotype is the product of probabilities across loci. This gives a multivariate discrete Laplace distribution, where the marginals (that is, at each locus) are independent, discrete Laplace distributions.

Just as before, for a one locus haplotype, we can assume that there is a modal/central Y-STR profile with r loci, $y = (y_1, y_2, \ldots, y_r)$, and all the alleles are distributed around this profile. We also assume that the discrete Laplace distribution at each locus has its own parameter, where p_k is the parameter at the k^{th} locus. Normally, the central Y-STR profile, y, would also be regarded as parameters.

As before, let $f(x; p, y)$ be the probability mass function of a discrete Laplace distribution. We define an observation $X = (X_1, X_2, \ldots, X_r)$ to be from a multivariate distribution of independent, discrete Laplace distributions when the probability of observing $X = x$ is

$$\prod_{k=1}^{r} f\left(x_k; p_k, y_k\right). \tag{1}$$

This corresponds to that the individual X has mutated away from y independently at each locus.

Now, we have one more generalisation. A population may have several subpopulations, e.g. introduced by migration or by evolution. This means that we need to have a mixture of multivariate distributions with marginally independent, discrete Laplace distributions. Each component in the mixture represents a subpopulation. We define an observation $X = (X_1, X_2, \ldots, X_r)$ to be from a mixture of multivariate, marginally independent, discrete Laplace distributions, when the probability of observing $X = x$ is

$$\sum_{j=1}^{c} \tau_j \prod_{k=1}^{r} f\left(x_k; p_{jk}, y_{jk}\right), \tag{2}$$

where τ_j is the a priori probability for originating from the j'th subpopulation. Thus, the parameters of this mixture model are $\{y_j\}_{j=1}^{c}$ with $y_j = (y_{j1}, y_{j2}, \ldots, y_{jr})$ as the central haplotype of the j^{th} subpopulation, $\{\tau_j\}_{j=1}^{c}$ and $\{p_{jk}\}_{\substack{j\in\{1,2,\ldots,c\}\\ k\in\{1,2,\ldots,r\}}}$ (the parameters for each discrete Laplace distribution).

We assume that p_{jk} depends on locus and subpopulation, such that $\log p_{jk} = \omega_j + \lambda_k$. This means that there is an additive effect of locus, λ_k, and an additive effect of subpopulation, ω_j.

More theory on finite mixture distributions is given by Titterington *et al.* (1987).

3.1. *Haplotype frequency prediction*

When we have estimated the parameters of a mixture of multivariate, marginally independent, discrete Laplace distributions (this will be shown in the next section), we can use these to estimate haplotype frequencies.

Given estimates of subpopulation central haplotypes $\{\hat{y}_j\}_j$, dispersion parameters $\{\hat{p}_{jk}\}_{j,k}$ and prior probabilities $\{\hat{\tau}_j\}_j$, the haplotype frequency of a haplotype $x = (x_1, x_2, \ldots, x_r)$ with $x_k \in \mathbb{Z}$ for $k \in \{1, 2, \ldots, r\}$ can be estimated as

$$\hat{p}(x) = \sum_{j=1}^{c} \hat{\tau}_j \prod_{k=1}^{r} f\left(x_k; \hat{p}_{jk}, \hat{y}_{jk}\right). \tag{3}$$

Thus, we simply use the estimated parameters in Equation (2) to obtain Equation (3).

4. Estimating parameters

In this section we demonstrate how to estimate the parameters in a mixture of multivariate, independent, discrete Laplace distributions. This can for example be used to estimate Y-STR haplotype frequencies.

First, the `R` package `disclapmix` (Andersen and Eriksen, 2013b; Andersen *et al.*, 2013) for analysing a mixture of multivariate, independent, discrete Laplace distributions must be loaded:

```
> library(disclapmix)
```

If you do not have the `disclapmix` package installed, please visit `http://cran.r-project.org/package=disclapmix`.

This package supplies the function `disclapmix` for estimating the parameters in a mixture of multivariate, marginally independent, discrete Laplace distributions with probability mass function given in Equation (2). We will refer to this as 'the discrete Laplace method'.

4.1. *Data from marginally independent, discrete Laplace distributions*

Now, we revisit the example leading to Figure 2 and add two more loci with different dispersion and location parameters. We then analyse the randomly generated values from independent, discrete Laplace distributions with a probability mass function as given in Equation (1).

```
> set.seed(1)
> n <- 100 # number of individuals
>
> # Locus 1
> p1 <- 0.3 # Dispersion parameter
> m1 <- 13 # Location parameter
> d1 <- rdisclap(n, p1) + m1 # Generate a sample
>
> # Locus 2
> p2 <- 0.4
```

```
> m2 <- 14
> d2 <- rdisclap(n, p2) + m2
>
> # Locus 3
> p3 <- 0.5
> m3 <- 15
> d3 <- rdisclap(n, p3) + m3
>
> db <- cbind(d1, d2, d3)
> db <- as.matrix(apply(db, 2, as.integer)) # Coerce to integer matrix
> head(db)
     d1 d2 d3
[1,] 14 15 16
[2,] 12 12 17
[3,] 13 13 15
[4,] 13 13 15
[5,] 14 12 15
[6,] 13 15 15
>
> # Fit the model (L means integer type)
> fit <- disclapmix(db, clusters = 1L)
```

We can then look at the estimated location parameters, $y = (y_1, y_2, y_3)$:

```
> fit$y
     d1 d2 d3
[1,] 13 14 15
```

And the estimated dispersion parameters, (p_1, p_2, p_3):

```
> fit$disclap_parameters
        d1     d2     d3
[1,] 0.265 0.4369 0.5167
```

As seen, the estimated dispersion location parameters are well estimated. The dispersion parameters are also quite close to the ones used to generate the data.

4.2. *Data from a Fisher-Wright population*

Andersen *et al.* (2013) simulated populations following the Fisher-Wright model of evolution (Fisher, 1922, 1930, 1958; Wright, 1931; Ewens, 2004) with assumptions of primarily neutral, single-step mutations of STRs (Ohta and Kimura, 1973). From these populations, data sets were sampled. Using the discrete Laplace method for estimating haplotype frequencies, the method worked rather well.

This is worth highlighting: Data was simulated under a completely different model than that used for inference afterwards. The data was simulated under a population model (Fisher-Wright model of evolution) with a certain mutation model (single-step mutation model). Inference was made assuming that the data was from a mixture of multivariate, marginally independent, discrete Laplace distributions.

One of the reasons that the discrete Laplace distribution predicts data from a Fisher-Wright model of evolution with a single-step mutation model is due to

the fact that it approximates certain properties of this population and mutation model (Caliebe *et al.*, 2010). This is also explained by Andersen *et al.* (2013).

Now, let us try simulating a Fisher-Wright population and analyse it with the discrete Laplace method. To simulate the population, the `R` package `fwsim` (Andersen and Eriksen, 2012b,a) is loaded:

```
> library(fwsim)
```

If you do not have the `fwsim` package installed, please visit `http://cran.r-project.org/package=fwsim`.

We then simulate a population consisting of Y-STR profiles:

```
> set.seed(1)
> generations <- 100
> population.size <- 1e+05
> number.of.loci <- 7
> mutation.rates <- seq(0.001, 0.01, length.out = number.of.loci)
> mutation.rates
[1] 0.0010 0.0025 0.0040 0.0055 0.0070 0.0085 0.0100
> sim <- fwsim(g = generations, k = population.size, r = number.of.loci,
>     mu = mutation.rates, trace = FALSE)
> pop <- sim$haplotypes
```

Note, that the mutation rates are different for each locus (ranging from 0.001 to 0.01). The location parameter is 0 for all loci by default. This can be changed afterwards without loosing or adding any information. Below, we change it to be $y = (14, 12, 28, 22, 10, 11, 13)$:

```
> y <- c(14, 12, 28, 22, 10, 11, 13)
> for (i in 1:number.of.loci) {
>     pop[, i] <- pop[, i] + y[i]
> }
> head(pop)
  Locus1 Locus2 Locus3 Locus4 Locus5 Locus6 Locus7 N
1     12     12     28     22     10     11     13 3
2     14     11     26     20      9     11     13 1
3     13     11     26     22     10     10     13 4
4     14     11     26     22      8     10     13 2
5     14     11     26     22      9     10     12 2
6     14     11     26     23     10     10     11 2
```

Then, y is the most frequent 10 locus Y-STR haplotype in Denmark according to `http://www.yhrd.org` (on March 26, 2013) restricted to the 7 loci minimal haplotype.

The column `N` is the number of individuals in the population with that Y-STR haplotype. Summing column `N` reveals that there is not exactly `population.size` individuals due to that the population size is stochastic (refer to Andersen and Eriksen (2012a) for the details).

We can then calculate the population frequency for each haplotype:

```
> pop$PopFreq <- pop$N/sum(pop$N)
```

Let us draw a data set where each haplotype is drawn relatively to its population frequency:

```
> set.seed(1)
> n <- 500 # Data set size
> types <- sample(x = 1:nrow(pop), size = n, replace = TRUE, prob = pop$N)
> types.table <- table(types)
>
> alpha <- sum(types.table == 1)
> alpha/n # Singleton proportion
[1] 0.492
> dataset <- pop[as.integer(names(types.table)), ]
> dataset$Ndb <- types.table
> head(dataset)
    Locus1 Locus2 Locus3 Locus4 Locus5 Locus6 Locus7   N   PopFreq Ndb
9       14     11     26     23     10      8     12   2 1.924e-05   1
103     14     11     28     19      9     10     12   1 9.619e-06   1
146     14     11     28     21     10     11     13 187 1.799e-03   3
229     14     11     27     21     11     12     12   6 5.771e-05   1
271     14     11     28     22      7     11     12  14 1.347e-04   1
273     14     11     28     22      8     11     12   6 5.771e-05   1
>
> db <- pop[types, 1:number.of.loci]
> db <- as.matrix(apply(db, 2, as.integer)) # Force integer matrix
> head(db)
     Locus1 Locus2 Locus3 Locus4 Locus5 Locus6 Locus7
[1,]     13     12     30     22      8     11     11
[2,]     14     12     28     22     10     11     14
[3,]     14     13     28     21     10     10     14
[4,]     14     12     28     22      9     11     14
[5,]     14     12     28     22     11     11     14
[6,]     14     12     28     22      9     10     14
```

Then, analyse it:

```
> fit <- disclapmix(db, clusters = 1L)
>
> # Estimated location parameters
> fit$y
     Locus1 Locus2 Locus3 Locus4 Locus5 Locus6 Locus7
[1,]     14     12     28     22     10     11     13
>
> # Estimated dispersion parameters
> fit$disclap_parameters
     Locus1 Locus2 Locus3 Locus4 Locus5 Locus6 Locus7
[1,] 0.0469  0.126 0.1589 0.1827 0.2453 0.2817  0.316
```

Let us compare the mutation rates with the dispersion parameters in the discrete Laplace distributions:

```
> plot(mutation.rates, fit$disclap_parameters, xlab = "Mutation rate",
>     ylab = "Estimated dispersion parameter")
```

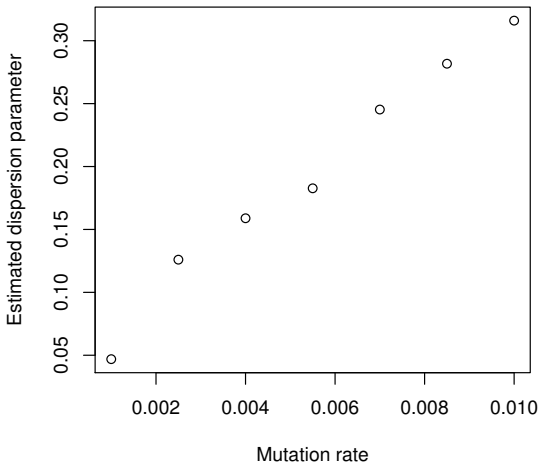


Figure 3. The relationship between the mutation rate in a Fisher-Wright population and the estimated dispersion parameters using the discrete Laplace method.

As expected, there is a connection between the mutation rate and the dispersion parameter (the exact connection is not known).

It is possible to predict a population frequency with the `predict` function as shown in Equation (3). This can be used to see how well the population frequency is predicted for each unique haplotype in the dataset (obtained by using `dataset` instead of `db`):

```
> pred.popfreqs <- predict(fit,
>   newdata = as.matrix(apply(dataset[, 1:number.of.loci], 2, as.integer)))
> plot(dataset$PopFreq, pred.popfreqs, log = "xy",
>   xlab = "True population frequency",
>   ylab = "Estimated population frequency")
> abline(a = 0, b = 1, lty = 1)
> legend("bottomright", "y = x (predicted = true)", lty = 1)
```

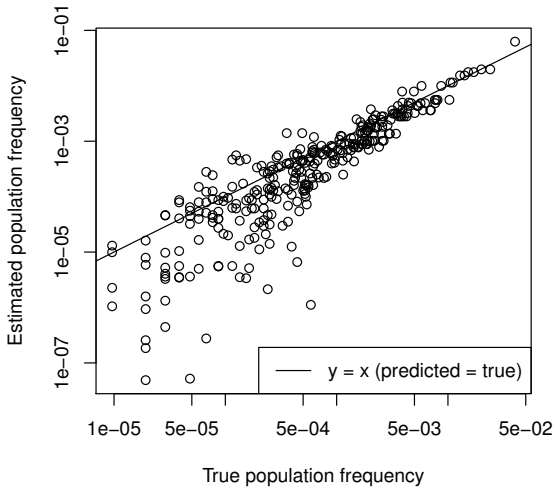


Figure 4. The relationship between the true population frequency and the predicted population frequency using the discrete Laplace method.

4.3. *Data from a mixture of two Fisher-Wright populations*

Here, we show how to analyse a dataset from a mixture of two populations. First, we simulate two populations (note the different mutation rates and location parameters, where the location parameters again are changed afterwards without loosing or adding any information):

```
> set.seed(1)
>
> # Common parameters
> generations <- 100
> population.size <- 1e+05
> number.of.loci <- 7
>
> mu1 <- seq(0.001, 0.005, length.out = number.of.loci)
> sim1 <- fwsim(g = generations, k = population.size, r = number.of.loci,
>     mu = mu1, trace = FALSE)
> pop1 <- sim1$haplotypes
> y1 <- c(14, 12, 28, 22, 10, 11, 13)
> for (i in 1:number.of.loci) pop1[, i] <- pop1[, i] + y1[i]
>
> mu2 <- seq(0.005, 0.01, length.out = number.of.loci)
> sim2 <- fwsim(g = generations, k = population.size, r = number.of.loci,
>     mu = mu2, trace = FALSE)
> pop2 <- sim2$haplotypes
> y2 <- c(14, 13, 29, 23, 11, 13, 13)
> for (i in 1:number.of.loci) pop2[, i] <- pop2[, i] + y2[i]
```

Here, just as $y_1 = (14,12,28,22,10,11,13)$ are the alleles from most frequent haplotype, then $y_2 = (14,13,29,23,11,13,13)$ are the alleles from the second most frequent haplotype.

Then we sample a data set with an expected proportion of 20% from the first population and 80% from the second population:

```
> set.seed(1)
> n <- 500 # Data set size
>
> n1 <- rbinom(1, n, 0.2)
> c(n1, n1/n)
[1] 102.000   0.204
>
> n2 <- n - n1
> c(n2, n2/n)
[1] 398.000   0.796
>
> types1 <- sample(x = 1:nrow(pop1), size = n1,
>   replace = TRUE, prob = pop1$N)
> db1 <- pop1[types1, 1:number.of.loci]
>
> types2 <- sample(x = 1:nrow(pop2), size = n2,
>   replace = TRUE, prob = pop2$N)
> db2 <- pop2[types2, 1:number.of.loci]
>
> db <- rbind(db1, db2)
> db <- as.matrix(apply(db, 2, as.integer)) # Force integer matrix
>
> # Singleton proportion
> sum(table(apply(db, 1, paste, collapse = ";")) == 1)/n
[1] 0.672
```

Now, we analyse the data set trying 1 to 5 subpopulations. Afterwards, we analyse the optimal number of subpopulations using the BIC (Bayesian Information Criteria) by Schwarz (1978):

```
> fits <- lapply(1L:5L,
>   function(clusters) disclapmix(db, clusters = clusters))
```

The BIC values are:

```
> BIC <- sapply(fits, function(fit) fit$BIC_marginal)
> BIC
[1] 9487 8600 8646 8700 8748
```

The estimated parameters for this optimal number of subpopulations can be made available in `best.fit` as follows:

```
> best.fit <- fits[[which.min(BIC)]]
> best.fit
disclapmixfit from 500 observations on 7 loci with 2 clusters.
>
> # Estimated a priori probability of originating from each
> # subpopulation
> best.fit$tau
[1] 0.2126 0.7874
>
> # Estimated location parameters
```

```
> best.fit$y
     Locus1 Locus2 Locus3 Locus4 Locus5 Locus6 Locus7
[1,]     14     12     28     22     10     11     13
[2,]     14     13     29     23     11     13     13
>
> # Estimated dispersion parameters for each subpopulation
> best.fit$disclap_parameters
         Locus1 Locus2 Locus3 Locus4 Locus5 Locus6 Locus7
cluster1 0.1029 0.1083 0.1213 0.1353 0.1458 0.1587 0.1595
cluster2 0.1896 0.1997 0.2234 0.2494 0.2686 0.2924 0.2938
```

The estimated location parameters are the same as those used for generating the data. Also, the values of τ_j, the a priori probability of originating from the j^{th} subpopulation, are consistent with the mixture proportions of 0.204 and 0.796.

We can also calculate the predicted population frequencies (using the mixture proportions 0.204 and 0.796):

```
> pop1$PopFreq <- pop1$N/sum(pop1$N)
> pop2$PopFreq <- pop2$N/sum(pop2$N)
>
> types1.table <- table(types1)
> types2.table <- table(types2)
>
> dataset1 <- pop1[as.integer(names(types1.table)), ]
> dataset1$Ndb <- types1.table
> sum(dataset1$Ndb)
[1] 102
>
> dataset2 <- pop2[as.integer(names(types2.table)), ]
> dataset2$Ndb <- types2.table
> sum(dataset2$Ndb)
[1] 398
>
> dataset <- merge(x = dataset1, y = dataset2,
>   by = colnames(db), all = TRUE)
> dataset[is.na(dataset)] <- 0
>
> dataset$MixPopFreq <- (n1/n)*dataset$PopFreq.x + (n2/n)*dataset$PopFreq.y
>
> dataset$Type <- "Only from pop1"
> dataset$Type[dataset$Ndb.y > 0] <- "Only from pop2"
> dataset$Type[dataset$Ndb.x > 0 & dataset$Ndb.y > 0] <- "Occurred in both"
> dataset$Type <- factor(dataset$Type)
```

We can now compare the predicted frequencies with the population frequency:

```
> pred.popfreqs <- predict(best.fit,
>   newdata = as.matrix(apply(dataset[, 1:number.of.loci], 2, as.integer)))
> plot(dataset$MixPopFreq, pred.popfreqs, log = "xy", col = dataset$Type,
>   xlab = "True population frequency",
>   ylab = "Estimated population frequency")
> abline(a = 0, b = 1, lty = 1)
> legend("bottomright",
>   c("y = x (predicted = true)", levels(dataset$Type)),
>   lty = c(1, rep(-1, 3)),
>   col = c("black", 1:length(levels(dataset$Type))),
>   pch = c(-1, rep(1, 3)))
```

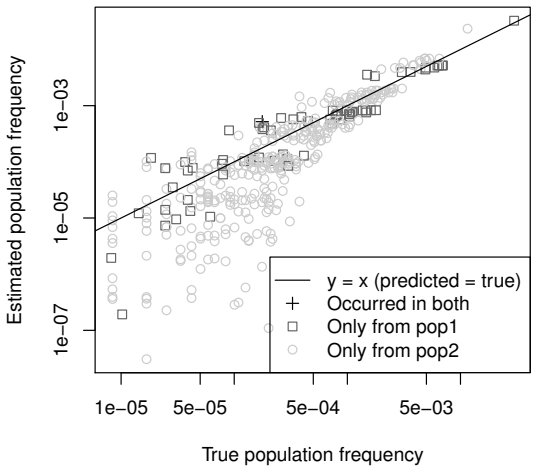


Figure 5. The relationship between the true population frequency and the predicted population frequency using the discrete Laplace method.

5. Concluding remarks

We have shown how to analyse Y-STR population data using the discrete Laplace method described by Andersen *et al.* (2013). This was done using the freely available and open source `R` packages `disclap`, `fwsim` and `disclapmix` that are supported on Linux, MacOS and MS Windows.

One key point made is worth repeating: Data simulated under a population model (e.g. the Fisher-Wright model of evolution) with a certain mutation model (e.g. the single-step mutation model) can be successfully analysed using the discrete Laplace method making inference assuming that the data is from a mixture of multivariate, independent, discrete Laplace distributions.

6. Bibliography

Andersen, M. M. and Eriksen, P. S. (2012a) Efficient Forward Simulation of Fisher-Wright Populations with Stochastic Population Size and Neutral Single Step Mutations in Haplotypes. *Preprint, arXiv:1210.1773*. 114

Andersen, M. M. and Eriksen, P. S. (2012b) *fwsim: Fisher-Wright Population Simulation*. URL `http://CRAN.R-project.org/package=fwsim`. R package version 0.2-5. 114

Andersen, M. M. and Eriksen, P. S. (2013a) *disclap: Discrete Laplace Family*. URL `http://CRAN.R-project.org/package=disclap`. R package version 1.4. 108

Andersen, M. M. and Eriksen, P. S. (2013b) *disclapmix: Discrete Laplace mixture inference using the EM algorithm*. URL `http://CRAN.R-project.org/package=disclapmix`. R package version 1.2. 112

Andersen, M. M., Eriksen, P. S. and Morling, N. (2013) The discrete Laplace exponential family and estimation of Y-STR haplotype frequencies. *Journal of Theoretical Biology*, **329**, 39–51. 108, 109, 111, 112, 113, 114, 120

Caliebe, A., Jochens, A., Krawczak, M. and Rösler, U. (2010) A Markov Chain Description of the Stepwise Mutation Model: Local and Global Behaviour of the Allele Process. *Journal of Theoretical Biology*, **266**, 336–342. 114

Ewens, W. J. (2004) *Mathematical Population Genetics*. Springer-Verlag. 113

Fisher, R. A. (1922) On the Dominance Ratio. *Proc. Roy. Soc. Edin.*, **42**, 321–341. 113

Fisher, R. A. (1930) *The Genetical Theory of Natural Selection*. Oxford: Clarendon Press. 113

Fisher, R. A. (1958) *The Genetical Theory of Natural Selection*. New York: Dover, 2nd revised edn. 113

Ohta, T. and Kimura, M. (1973) A Model of Mutation Appropriate to Estimate the Number of Electrophoretically Detectable Alleles in a Finite Population. *Genet. Res.*, **22**, 201–204. 113

R Development Core Team (2013) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org`. ISBN 3-900051-07-0. 108

Schwarz, G. (1978) Estimating the Dimension of a Model. *Annals of Statistics*, **6**, 461–464. 118

Titterington, D. M., Smith, A. F. M. and Makov, U. E. (1987) *Statistical Analysis of Finite Mixture Distributions*. Wiley. 111

Wright, S. (1931) Evolution in Mendelian populations. *Genetics*, **16**, 97–159. 113

Paper VIII

Cluster analysis of European Y-chromosomal STR haplotypes using the discrete Laplace method

Author list Mikkel Meyer Andersen, *Aalborg University, Denmark*
Poul Svante Eriksen, *Aalborg University, Denmark*
Niels Morling, *University of Copenhagen, Denmark*

Summary The European Y-chromosomal short tandem repeat (STR) haplotype distribution has previously been analysed in various ways. Here, we introduce a new way of analysing population substructure using a new method based on clustering within the discrete Laplace exponential family that models the probability distribution of the Y-STR haplotypes. Creating a consistent statistical model of the haplotypes in a probability distribution framework enables us to perform a wide range of analyses. A very important practical fact is that the calculations can be performed on a normal computer.

We identified two sub-clusters of the Eastern and Western European Y-STR haplotypes similar to results of previous studies. We also compared pairwise distances (between geographically separated samples) with those obtained using the AMOVA method and found good agreement. Furthermore, we investigated the homogeneity in two different ways and found that the Y-STR haplotypes from e.g. Finland were relatively homogeneous as opposed to the relatively heterogeneous Y-STR haplotypes from e.g. Lublin, Eastern Poland and Berlin, Germany. We demonstrated that the observed distributions of alleles at each locus were similar to the expected ones.

Publication info This paper has been submitted to Forensic Science International: Genetics (2013).

1. Introduction

Recent historical events in the European Y-chromosomal short tandem repeat (Y-STR) haplotype distribution were analysed by Roewer *et al.* (2005) based upon a database with approximately 12,700 Y-STR profiles from 91 different locations in Europe. The analysis was performed by means of AMOVA (Excoffier *et al.*, 1992), which is a cluster analysis method based upon molecular variance. In this paper, we analysed the same data using a new method based on a combination of multivariate, marginally independent, discrete Laplace distributions (called 'discrete Laplace method') as described by Andersen *et al.* (2013b) and practically introduced by Andersen *et al.* (2013a). We demonstrate how to use the discrete Laplace method for making inference in Y-STR haplotype databases.

The AMOVA method (Excoffier *et al.*, 1992) is widely used in population and forensic genetics. The AMOVA method introduced the molecular variance measure Φ_{ST} that is an analogue to Wright's F_{ST}. Φ_{ST} is based on the minimum detectable evolutionary distances between individual haplotypes. When a population consists of different strata (for example geographically separated sampling locations), AMOVA can be used to infer stratification through non-parametric cluster analysis of the Φ_{ST} distances.

Whereas the AMOVA method performs non-parametric cluster analysis of the Φ_{ST} distances, the discrete Laplace method by Andersen *et al.* (2013b) models the probability distribution of the Y-STR haplotypes. This makes it possible to perform much more detailed inference, e.g. estimating haplotype frequencies, model based cluster analysis, analysis of population homogeneity and comparing the observed distribution of alleles at each locus to the expected one. We note that the calculations can be performed on a normal computer.

2. Method

Assume that we have S different strata (for example sample locations), each with n_s individuals for $s \in \{1,2,\ldots,S\}$, and that there are $n = \sum_{s=1}^{S} n_s$ individuals in total. Let $x_i = (x_{i1}, x_{i2}, \ldots, x_{ir})$ be the r loci Y-STR haplotype for the i'th individual for $i \in \{1,2,\ldots,n\}$. Let I_s be the indices for the individuals in the s'th stratum.

We now assume that the parameters in the discrete Laplace method (Andersen *et al.*, 2013b) are estimated, for example by using the `R` (R Development Core Team, 2013) library `disclapmix` version 1.2 (Andersen and Eriksen, 2013) that are described and demonstrated with both simple and more advanced examples in Andersen *et al.* (2013a). The estimated parameters are:

- The number of components in the mixture that can be interpreted as the number of estimated (genetic) subpopulations, $\hat{c}$ (from now on just c for easier notation).

- The central haplotype, $\hat{y}_j = (\hat{y}_{j1}, \hat{y}_{j2}, \ldots, \hat{y}_{jr})$, of the subpopulations for $j \in \{1,2,\ldots,c\}$. Subpopulations are constructed such that the individuals are close to the central haplotype.

- The prior probabilities, $\hat{\tau}_j$ for $j \in \{1,2,\ldots,c\}$, of belonging to the j'th subpopulation.
- The parameters of the multivariate, marginally independent, discrete Laplace distributions, $\hat{p}_{jk} = \exp(\hat{\omega}_j + \hat{\lambda}_k)$ for $j \in \{1,2,\ldots,c\}$ and $k \in \{1,2,\ldots,r\}$. This means that there is an additive effect of locus, $\hat{\lambda}_k$, and an additive effect of subpopulation, $\hat{\omega}_j$, as described by Andersen *et al.* (2013b).

The subpopulation membership can be formulated as

$$v_{ij} = \begin{cases} 1 & \text{if the } i\text{'th individual originates from the } j\text{'th subpopulation,} \\ 0 & \text{otherwise.} \end{cases}$$

Thus, an individual can originate from only one subpopulation. Because the membership is not observed, the probability of each outcome is instead estimated using the EM algorithm by Dempster *et al.* (1977) as described by Andersen *et al.* (2013b). Thus, given x_i, let $\hat{v}_{ij}$ be the estimated probability that the i'th individual originates from the j'th subpopulation. Thus, $\hat{v}_{i+} = \sum_{j=1}^{c} \hat{v}_{ij} = 1$ for all i. The estimation procedure described by Andersen *et al.* (2013b) results in

$$\hat{v}_{+j} = \sum_{i=1}^{n} \hat{v}_{ij} = \hat{\tau}_j$$

being the prior probability of belonging to the j'th subpopulation.

The estimate of the parameter, c, the number of subpopulations, can be obtained by using e.g. the Bayesian information criteria (BIC) (Schwarz, 1978) for various choices of c.

3. Analysis

The dataset analysed is a European 7-loci Y-STR database from 2004 consisting of 12,727 individuals in 91 strata (European sample locations). This dataset was first analysed by Roewer *et al.* (2005) using AMOVA (Excoffier *et al.*, 1992) among other analysis methods. The 7 Y-STR loci were `DYS19`, `DYS389I`, `DYS389II`, `DYS390`, `DYS391`, `DYS392` and `DYS393`. The alleles at `DYS389II` were replaced by `DYS389II` minus `DYS389I` (Butler, 2005).

For parameter estimation using the discrete Laplace method, 40 subpopulations were found to be optimal among the subpopulation counts that we used (which were from 5 to 60 at intervals of 5). This was done using the `disclapmix` library version 1.2 for the statistical software `R` (R Development Core Team, 2013) as shown below:

```
> library(disclapmix)
> str(db) # Note, the db must be an integer matrix
 int [1:12727, 1:7] 12 12 13 13 13 13 13 13 13 13 ...
 - attr(*, "dimnames")=List of 2
  ..$ : NULL
  ..$ : chr [1:7] "DYS19" "DYS389I" "DYS389II" "DYS390" ...
> head(db)
```

```
     DYS19 DYS389I DYS389II DYS390 DYS391 DYS392 DYS393
[1,]    12      13       17     24     10     11     13
[2,]    12      13       17     24     10     11     14
[3,]    13      12       18     24     10     11     13
[4,]    13      13       16     23     10     11     13
[5,]    13      13       16     24     10     11     14
[6,]    13      13       16     24     11     13     13
> head(popnames)
[1] Albania Albania Albania Albania Albania Albania
91 Levels: Albania Anatolia,Turkey Andalusia,Southern_Spain ... Zeeland,
    South-Western_Netherlands
> fits <- lapply(seq(5L, 60L, 5), function(clusters) disclapmix(db,
    clusters = clusters))
> fits_BIC <- sapply(fits, function(fit) fit$BIC_marginal)
> bestfit <- fits[[which.min(fits_BIC)]]
> summary(bestfit)
```

Please, see the values of the marginal BICs in Table 1. From now on, we focus on the results from the model with 40 subpopulations as this subset gave the best BIC score.

Subpopulations	BIC value
5	196,524.9
10	187,973.4
15	183,594.7
20	182,215.6
25	181,407.6
30	180,645.7
35	180,531.6
40	180,524.8
45	180,582.8
50	180,555.7
55	180,551.9
60	180,735.2

Table 1. The values of the marginal BICs at subpopulation counts from 5 to 60 at intervals of 5.

In Figure 1, the number of times that a haplotype was observed was compared to the estimated haplotype frequency using the discrete Laplace method. Haplotype frequency estimation using the discrete Laplace method was performed as follows: Given the central haplotype of the subpopulations, $\hat{y}_j$ for $j \in \{1,2,\ldots,c\}$, parameters $\hat{p}_{jk}$ for $j \in \{1,2,\ldots,c\}$ and $k \in \{1,2,\ldots,r\}$ and prior probabilities $\hat{\tau}_j$ for $j \in \{1,2,\ldots,c\}$ (here from the fitted model with 40 subpopulations), the haplotype frequency of a haplotype $h = (h_1, h_2, \ldots, h_r)$ with $h_k \in \mathbb{Z}$ for $k \in \{1,2,\ldots,r\}$ was estimated as

$$\sum_{j=1}^{c} \hat{\tau}_j \prod_{k=1}^{r} f\left(|h_k - \hat{y}_{jk}|; \hat{p}_{jk}\right),$$

where

$$f(d;p) = \left(\frac{1-p}{1+p}\right) p^{|d|}$$

is the probability mass function of a discrete Laplace distribution with parameter $0 < p < 1$ evaluated at $d \in \mathbb{Z}$.

This can be done using the `disclapmix` library for all haplotypes in the dataset:

```
> estimates <- predict(bestfit, newdata = db)
```

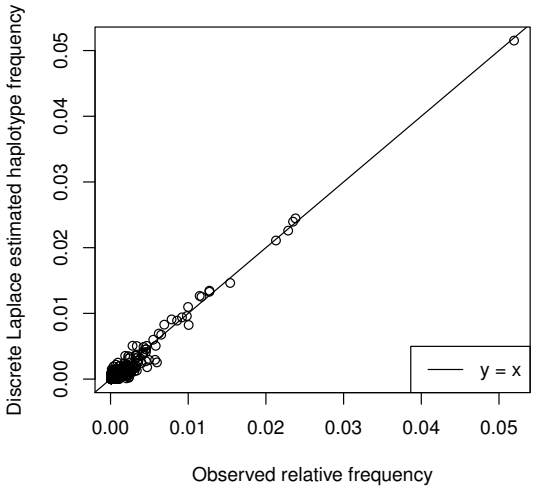


Figure 1. Comparison of (1) the relative frequency of a haplotype (number of times it has been observed dividided by database size) (2) the estimated haplotype frequency using the discrete Laplace method. Note, that for frequently observed haplotypes, the estimated haplotype frequency using the discrete Laplace method is close to the relative frequency.

3.1. *Model based cluster analysis*

As already mentioned, given the i'th individual's haplotype, x_i, let $\hat{v}_{ij}$ denote the estimated probability that the i'th individual originates from the j'th subpopulation. In this section, we analyse the $\hat{v}_{ij}$ values in a number of different ways.

To measure a distance between two subpopulations, a naïve approach of taking the minimum number of mutations between the central haplotype of the subpopulations, $\hat{y}_j$, was initially tried. Because this resulted in a large number of ties, a more sophisticated method based on the symmetrized Kullback-Leibler divergence (using the discrete Laplace method) was used. This distance measure is described in Appendix A. The distance between two subpopulations, j_1 and j_2, is denoted by

$$\mathrm{KL}(j_1, j_2). \tag{1}$$

Now, we have a distance measure between subpopulations, and we introduce a summary of the $\hat{v}_{ij}$ values for each stratum, s, and each subpopulation, j. Let

$$w_{sj} = n_s^{-1} \sum_{i \in I_s} \hat{v}_{ij} \tag{2}$$

be the s'th stratum's mean probability of originating from the j'th subpopulation for $s \in \{1,2,\ldots,S\}$ and $j \in \{1,2,\ldots,c\}$. Note, that

$$w_{s+} = \sum_{j=1}^{c} w_{sj} = 1 \quad \text{and} \quad w_{+j} = \sum_{s=1}^{S} w_{sj} = \hat{\tau}_j.$$

As mentioned, 40 subpopulations were found to be the optimal number using BIC. In Figure 2, a map of Europe with the w_{sj} values for all subpopulations $j \in \{1,2,\ldots,c\}$ at each stratum, s (sampling locations), is shown. The values were calculated as shown below:

```
> vij <- bestfit$v_matrix
> wsj <- aggregate(vij, list(popnames), mean)
> rownames(wsj) <- wsj[, 1]
> wsj <- as.matrix(wsj[, -1])
```

The majority of the central haplotypes of the subpopulations were close to each other. To better visualise the subpopulations, those with central haplotypes close to each other were assembled into mega clusters. Given a desired number of clusters, T, let J_t for $t \in \{1,2,\ldots,T\}$ be a partition of $\{1,2,\ldots,c\}$ such that

$$\bigcup_{t=1}^{T} J_t = \{1,2,\ldots,c\} \quad \text{and} \quad \bigcap_{t=1}^{T} J_t = \emptyset,$$

where $\emptyset = \{\}$ is the empty set. The collapsed w_{sj} values are

$$u_{st} = \sum_{j \in J_t} w_{sj},$$

such that

$$u_{+t} = \sum_{s=1}^{S} u_{st} = \sum_{s=1}^{S} \sum_{j \in J_t} w_{sj} = \sum_{j \in J_t} \sum_{s=1}^{S} w_{sj} = \sum_{j \in J_t} \hat{\tau}_j$$

for the t'th cluster and

$$u_{s+} = \sum_{t=1}^{T} u_{st} = \sum_{t=1}^{T} \sum_{j \in J_t} w_{sj} = 1$$

for the s'th stratum.

This means that we add together subpopulations J_t by adding their respective w_{sj} values for $j \in J_t$ to obtain mega clusters. Note, that the strata (or information about strata) are not used for constructing the mega clusters; only the central haplotype of the subpopulations and $\hat{p}_{jk}$ parameters are used.

Motivated by Roewer *et al.* (2005), two mega clusters were made based on the $\mathrm{KL}(j_1, j_2)$ distances between the central haplotype of subpopulations. Looking at the resulting u_{st} values on a European map as shown in Figure 3, it seems as if an Eastern and a Western European population emerge.

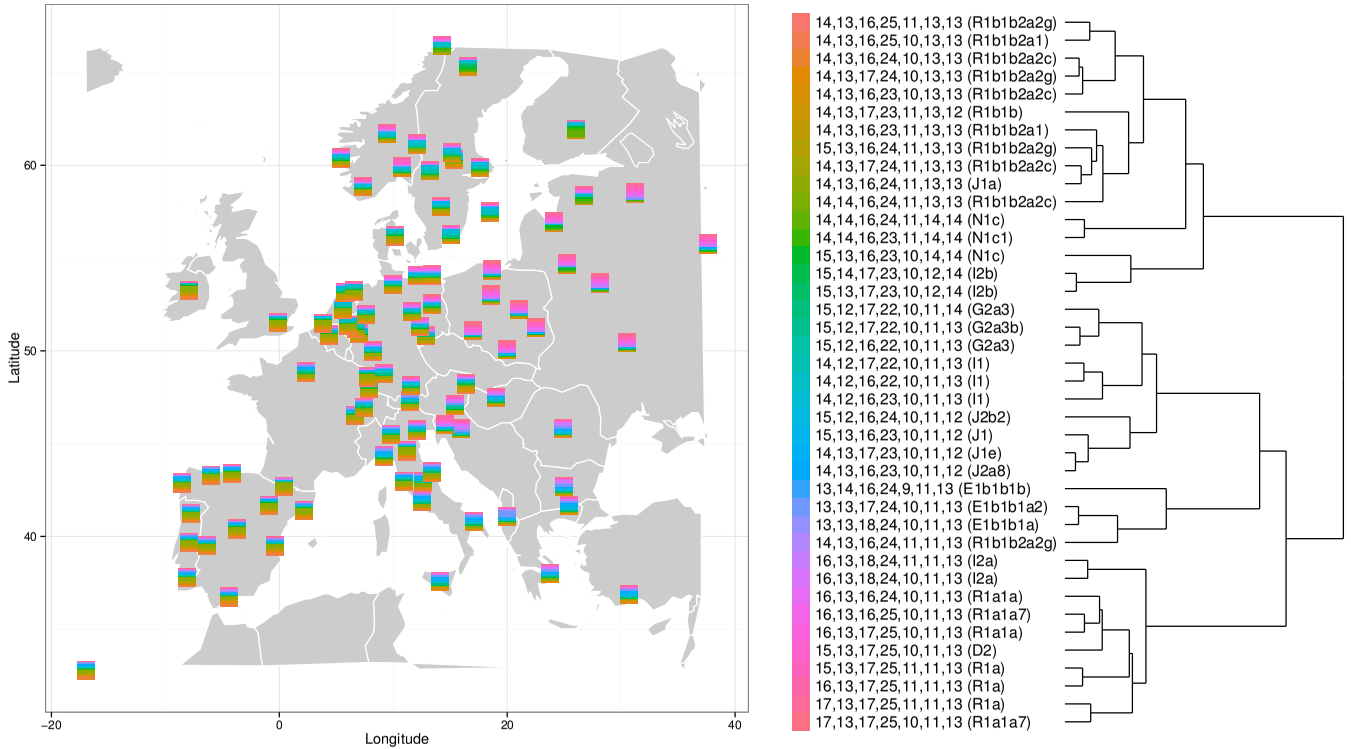


Figure 2. Map of Europe with the w_{sj} values for all subpopulations $j \in \{1,2,\ldots,c\}$ at each stratum s (sampling locations). A dendogram based on complete hierachical clustering (Sørensen, 1948; Defays, 1977) (such that the distance between two clusters is the maximum distance between their individual haplotypes) of the central haplotype y_j of subpopulation for $j \in \{1,2,\ldots,c\}$ using $\mathrm{KL}(j_1, j_2)$ given in Equation (1) as the distance metric is shown together with the colours for each of the $c = 40$ subpopulations. The leaves (subpopulations) of the dendogram have been reordered using the `R` (R Development Core Team, 2013) library `seriation` (Hahsler *et al.*, 2008, 2012) with the `OLO` method (Hahsler *et al.*, 2008). The labels are the central haplotype of the corresponding subpopulation and the predicted haplogroup from `http://www.yhrd.org` release 44 (Roewer *et al.*, 2001; Willuweit and Roewer, 2009). The 7 loci are (in order): `DYS19`, `DYS389I`, `DYS389II`, `DYS390`, `DYS391`, `DYS392` and `DYS393`, where the alleles at `DYS389II` were replaced by `DYS389II` minus `DYS389I` (Butler, 2005).

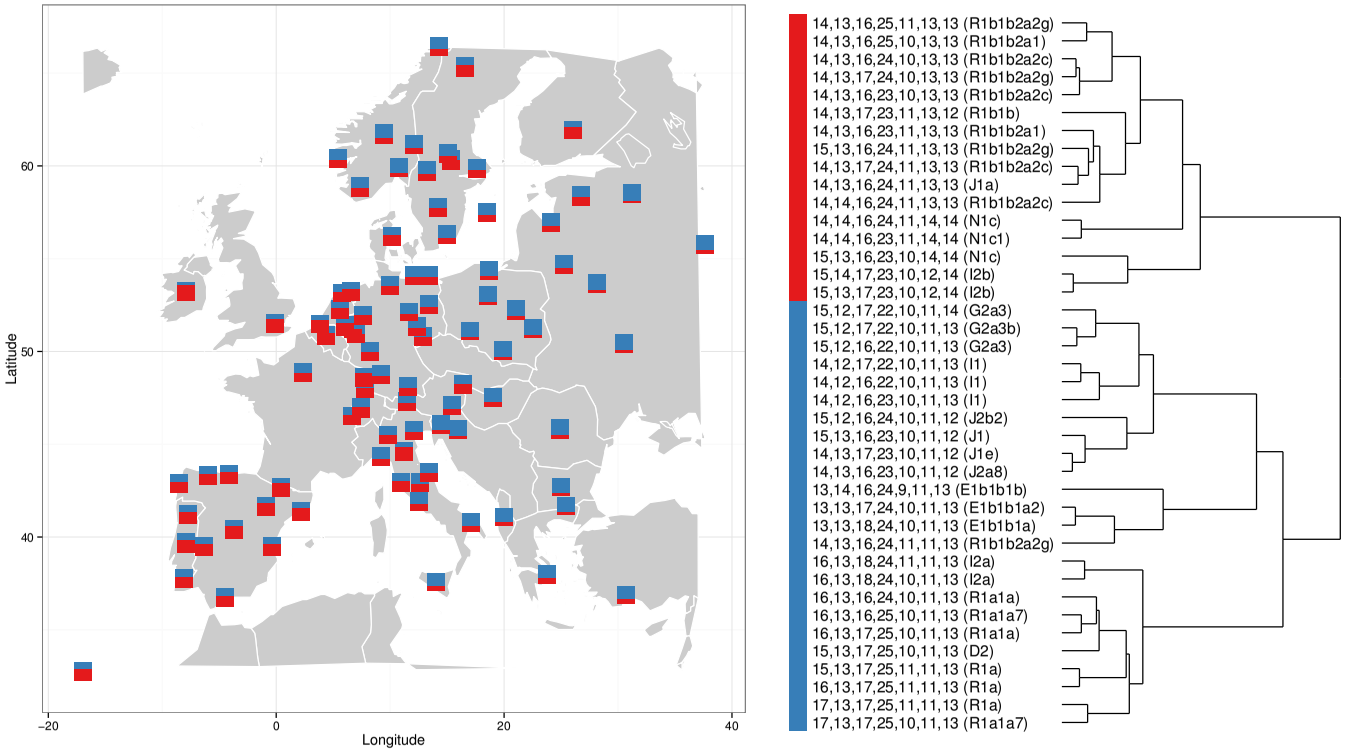


Figure 3. Map of Europe with the u_{st} values for all clusters $t \in \{1,2\}$ at each stratum s (sampling locations). Please refer to the caption of Figure 2 for a description of the hierachical clustering and the labels of the central haplotypes. The partitions are indicated by colours.

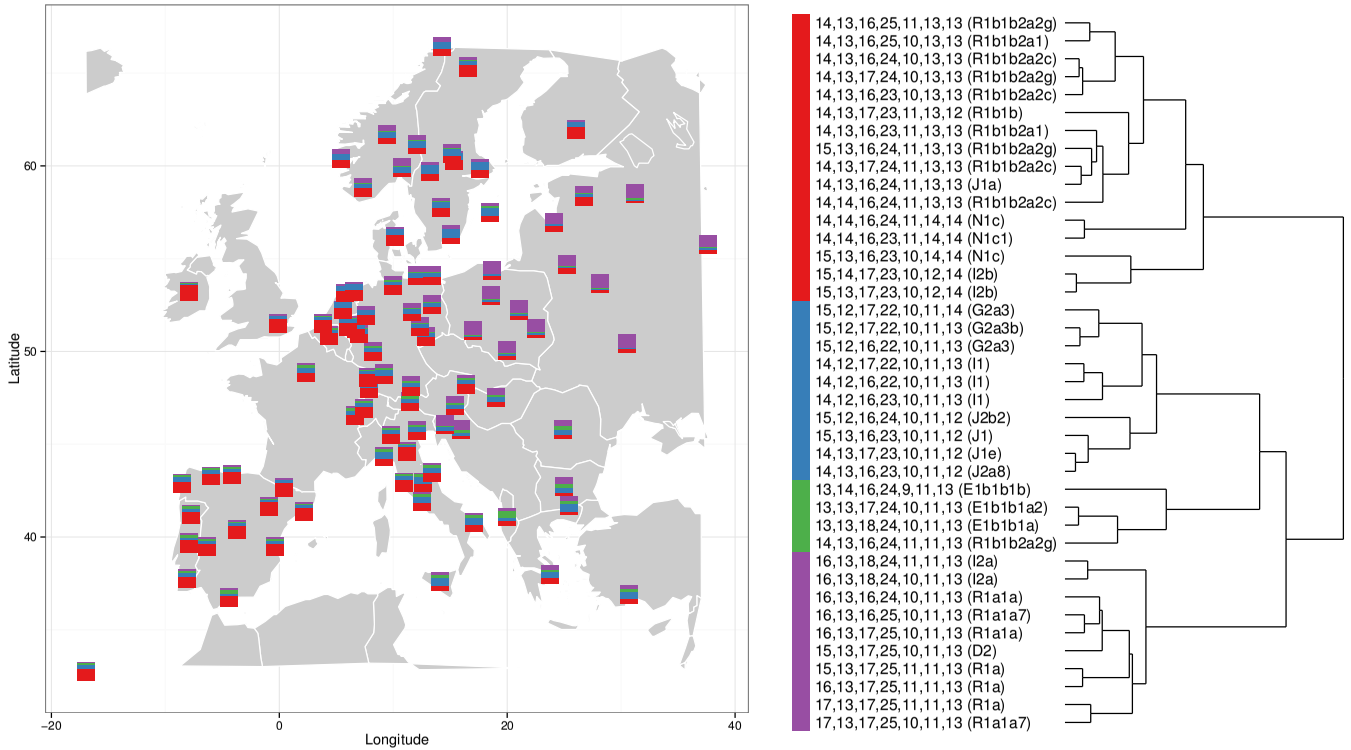


Figure 4. Map of Europe with the u_{st} values for all clusters $t \in \{1,2,3,4\}$ at each stratum s (sampling locations). Please refer to the caption of Figure 2 for a description of the hierachical clustering and the labels of the central haplotypes. The partitions are indicated by colours.

If four mega clusters were chosen, a map as shown in Figure 4 was obtained. As seen, it was now possible to identify Northern (Scandinavia), Southern (near the Balkan Peninsula), Eastern and Western European populations.

Pairwise distances

Let

$$\delta(s,t) = \sum_{j=1}^{c} \left(w_{sj} - w_{tj}\right)^2 \tag{3}$$

be the pairwise (L_2) distance between stratum s and stratum t using the mean estimated subpopulation affiliations w_{sj} and w_{tj} introduced in Equation (2). This is the squared Euclidean distance between vector $(w_{s1}, w_{s2}, \ldots, w_{sr})$ and vector $(w_{t1}, w_{t2}, \ldots, w_{tr})$. This can for example be used for hierachical clustering, as seen in Figure 6. For comparison, see Figure 7 for hierachical clustering of the pairwise Φ_{ST} distances calculated with Arlequin version 3.5 (Excoffier and Lischer, 2010) that uses the AMOVA method by Excoffier *et al.* (1992).

These pairwise distances can be compared as shown in Figure 5. As seen, there is a strong correlation between the Φ_{ST} values and the $\delta(s,t)$ values even though they are calculated in two very different ways.

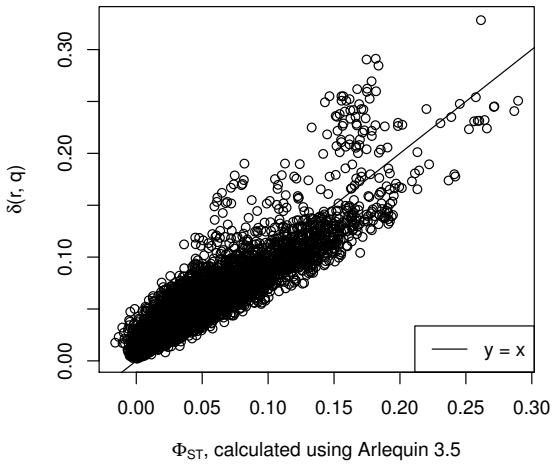


Figure 5. Comparision of Φ_{ST} distances (from the AMOVA method of Excoffier *et al.* (1992) calculated using Arlequin version 3.5 (Excoffier and Lischer, 2010)) and the $\delta(s,t)$ distances (calculated using the discrete Laplace method). Pearson's correlation coefficient: 0.90 with p-value $< 10^{-15}$.

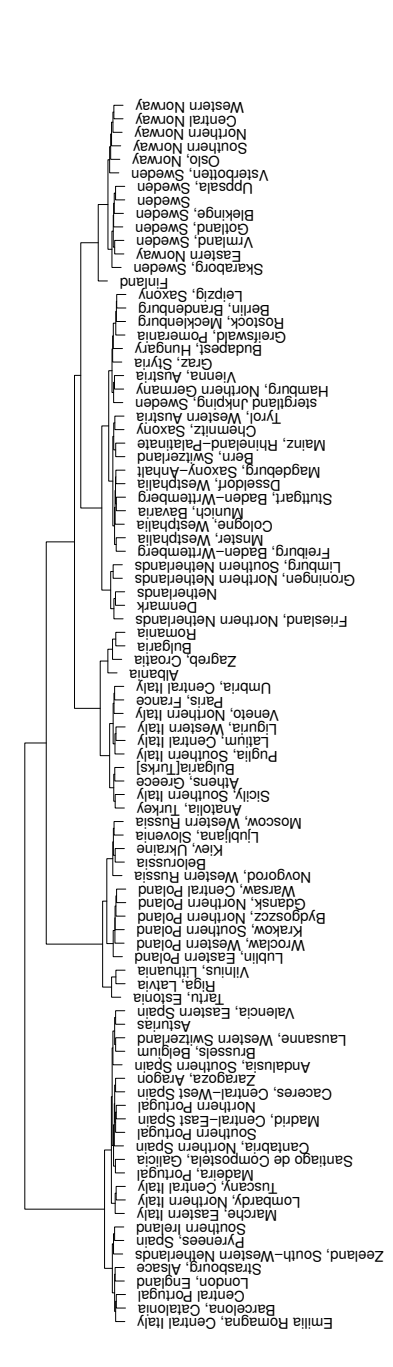


Figure 6. A dendogram of $\delta(s,t)$ distances (using the discrete Laplace method) between all pairs of strata as defined in Equation (3). The dendogram was made by hierachical clustering using Ward's method (Ward, 1963).

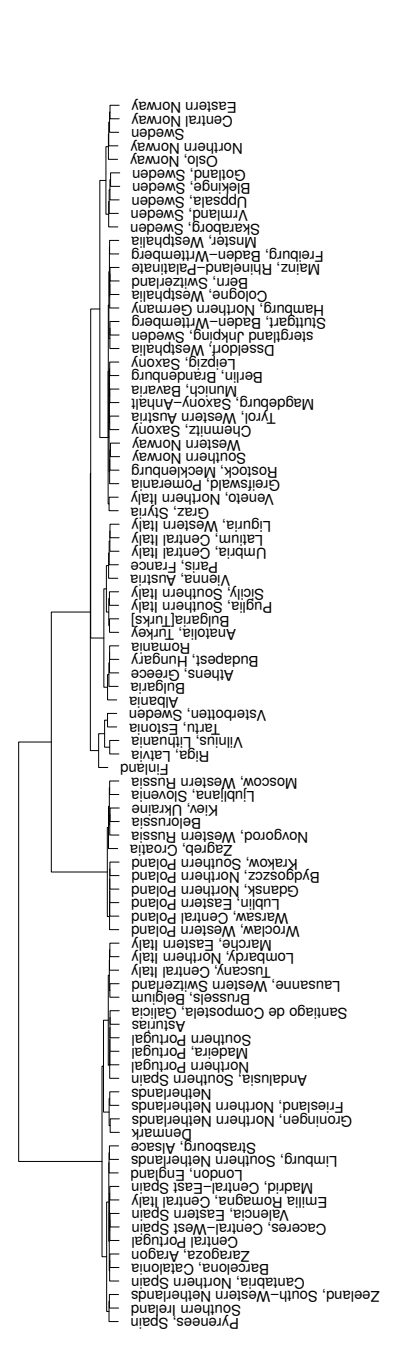


Figure 7. A dendogram of Φ_{ST} distances (Excoffier et al., 1992) between all pairs of strata calculated using Arlequin version 3.5 (Excoffier and Lischer, 2010). The dendogram was made by hierachical clustering using Ward's method (Ward, 1963).

Population homogeneity

In this section, we focus on two different homogeneity measures for strata and exemplify these measures by looking at three strata (sample locations).

First, let

$$H_s = \sum_{j=1}^{c} w_{sj} \log w_{sj}$$

be the homogeneity entropy of the s'th stratum for $s \in \{1,2,\ldots,S\}$.

Let

$$e_i = \sum_{j=1}^{c} \hat{v}_{ij} \log \hat{v}_{ij}$$

be the entropy of the i'th individual for $i \in \{1,2,\ldots,n\}$, and let

$$P_s = n_s^{-1} \sum_{i \in I_s} e_i$$

be the subpopulation certainty entropy of the s'th stratum for $s \in \{1,2,\ldots,S\}$.

Note, that H_s is the entropy of the $\hat{v}_{ij}$ means whereas P_s is the mean of the $\hat{v}_{ij}$ entropies.

Three extreme strata are now chosen for further investigations. These three strata are 'Finland' (lowest homogeneity entropy $H_s = 2.29$ and lowest subpopulation certainty entropy $P_s = 0.60$), 'Lublin, Eastern Poland' (homogeneity entropy $H_s = 3.07$ and highest subpopulation certainty entropy $P_s = 1.16$) and 'Berlin, Brandenburg, Germany' (highest homogeneity entropy $H_s = 3.43$ and subpopulation certainty entropy $P_s = 0.86$).

In Figure 8, $\hat{v}_{ij}$ values are plottet for 'Finland', 'Lublin, Eastern Poland' and 'Berlin, Brandenburg, Germany'. This can also be done for the four mega clusters and the result of this is shown in Figure 9.

3.2. *Marginals*

To validate a model of the Y-STR haplotype probability distribution, a reasonable validation criterium is that the predicted single and pairwise marginal allele distributions fit well with the observed distributions. This means that if 50% of the individuals in the database have allele 14 at `DYS19` (disregarding the alleles at the other loci), then this should also be predicted by the discrete Laplace method.

Single marginals

For each locus, the observed marginal distribution (percentage of individuals having each allele) can be compared with the expected distribution under the discrete Laplace method that is given by

$$P(x) = \sum_{j=1}^{c} \hat{\tau}_j f\left(|x - \hat{y}_{jk}|; \hat{p}_{jk}\right)$$

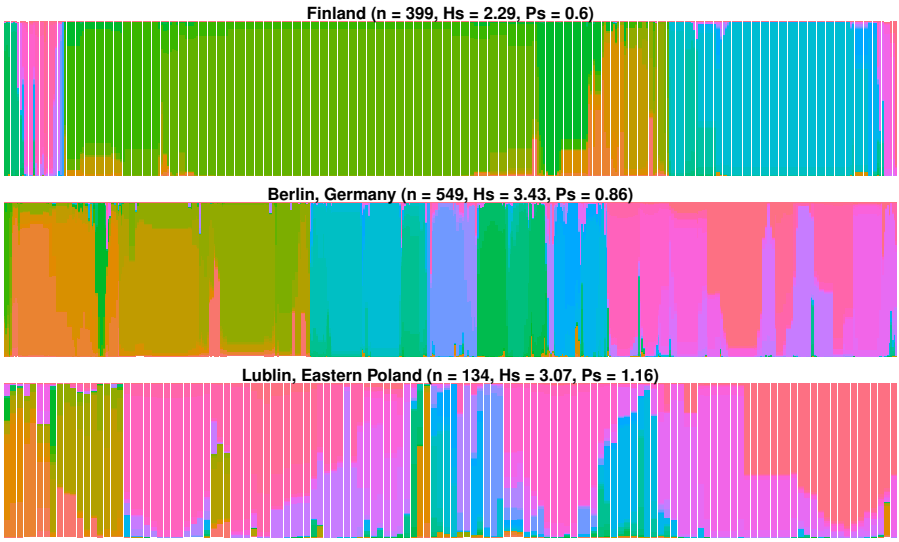


Figure 8. Each vertical bar shows an individual's row in the v_{ij} matrix (such that the i'th vertical bar consists of the 40 numbers $\{v_{ij}\}_j$ for $j \in \{1,2,\ldots,40\}$). The v_{ij} matrices are shown for Finland (lowest homogeneity entropy, $H_s = 2.29$, and lowest subpopulation certainty entropy, $P_s = 0.60$), Lublin, Eastern Poland (homogeneity entropy, $H_s = 3.07$, and highest subpopulation certainty entropy, $P_s = 1.16$) and Berlin, Germany (highest homogeneity entropy, $H_s = 3.43$, and subpopulation certainty entropy, $P_s = 0.86$). The subpopulations (the columns of the v_{ij} matrices) have the same order and colour as in Figure 2. In Figure 9, a similar figure is shown for four mega clusters. The individuals were reordered using the `R` library `seriation` (Hahsler et al., 2012, 2008) with the `BEA_TSP` method (Hahsler et al., 2008).

for each allele x at the k'th locus. Figure 10 shows the single marginal distribution for each locus. Note, that this is a mixture of discrete Laplace distributions, which means that it is not necessarily shaped like a single, discrete Laplace distribution.

Pairwise marginals

For two loci, k and l, the observed marginal distribution (number of individuals having each combination of alleles at the two loci) can be compared with the expected distribution under the discrete Laplace method that is given by

$$P(x_k, x_l) = \sum_{j=1}^{c} \hat{\tau}_j f\left(|x_k - \hat{y}_{jk}|; \hat{p}_{jk}\right) f\left(|x_l - \hat{y}_{jl}|; \hat{p}_{jl}\right)$$

for alleles (x_k, x_l) for locus k and l, respectively.

4. Discussion

We have demonstrated that the discrete Laplace method (analysing a mixture of multivariate, marginally independent, discrete Laplace distributions) as

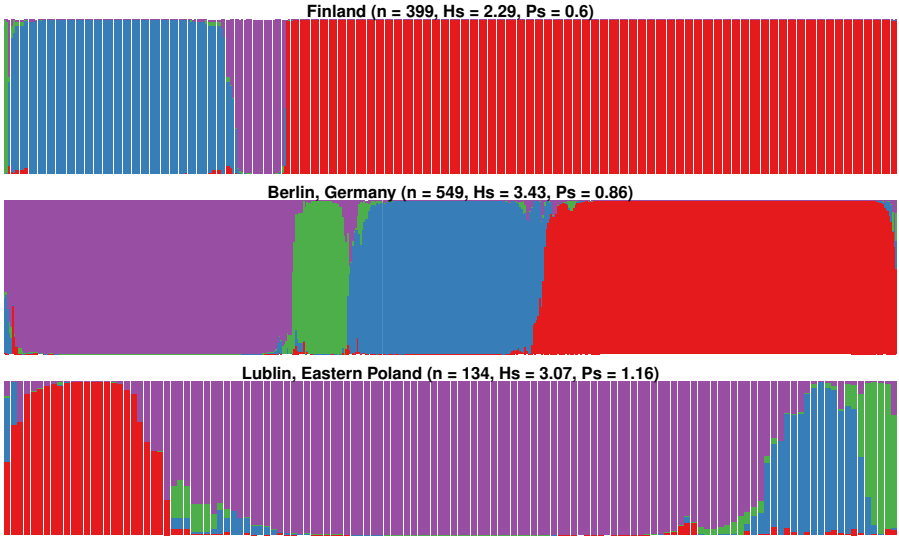


Figure 9. Please refer to the caption of Figure 8. The subpopulations (the columns of the merged v_{ij} matrices were reordered such that four mega clusters were obtained) have the same order and colour as in Figure 4.

described by Andersen *et al.* (2013b,a) is a valuable tool for modelling Y-chromosomal STR haplotypes and for making inference based on such a modelling. The discrete Laplace method can be used for a wide range of tasks such as haplotype frequency estimation and model based cluster analysis (e.g. in analysing substructure). Furthermore, the calculations can be performed on a normal computer.

In the model based cluster analysis performed in Section 3.1, Western and Eastern European subpopulations were identified (refer to Figure 3) similar to the results of Roewer *et al.* (2005) obtained using the AMOVA method by Excoffier *et al.* (1992). A more detailed map of Europe using all identified subpopulations is shown in Figure 2.

Another comparison of the discrete Laplace method with the AMOVA method (Excoffier *et al.*, 1992) was performed in Section 3.1. Here, it was shown that there was good agreement between the pairwise distances between strata (geographically separated sampling locations) obtained using the discrete Laplace method and the AMOVA method.

Homogeneity was analysed in two different ways, see Section 3.1. We found that the Y-STR haplotypes from Finland were more homogeneous than those from Lublin, Eastern Poland and Berlin, Germany (refer to Figure 8). Lublin is known to have been a center for trade (Lerski, 1996), so this heterogeneity seems quite reasonable.

The discrete Laplace method makes it possible to calculate the expected distribution of alleles (expected percentage of individuals having a certain allele). We demonstrated that the expected distribution of alleles at each locus was similar to the observed distribution (refer to Section 3.2 and Figure 10).

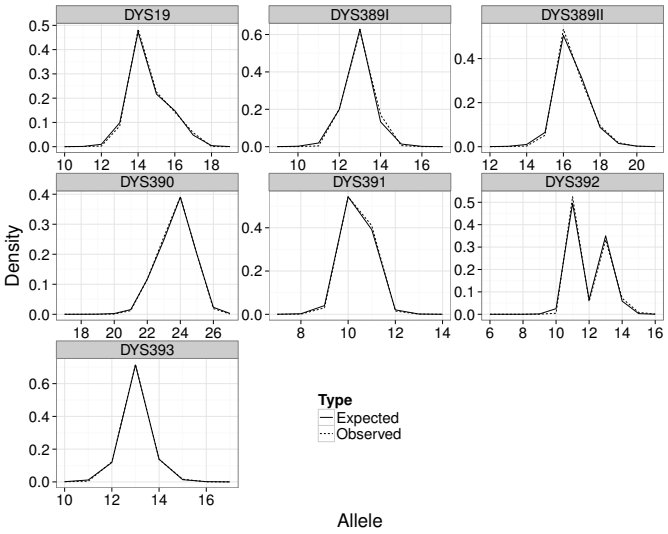


Figure 10. Single marginal observed and expected (by the discrete Laplace method) distributions for each Y-STR locus.

5. Acknowledgements

Thanks to both Oticon Foundation and Ellen and Aage Andersen's Foundation who supported parts of this work.

6. Bibliography

Andersen, M. M. and Eriksen, P. S. (2013) *disclapmix: Discrete Laplace mixture inference using the EM algorithm*. URL `http://CRAN.R-project.org/package=disclapmix`. R package version 1.2. 124

Andersen, M. M., Eriksen, P. S. and Morling, N. (2013a) A gentle introduction to the discrete Laplace method for estimating Y-STR haplotype frequencies. *Preprint, arXiv:1304.2129*. 124, 136

Andersen, M. M., Eriksen, P. S. and Morling, N. (2013b) The discrete Laplace exponential family and estimation of Y-STR haplotype frequencies. *Journal of Theoretical Biology*, **329**, 39–51. 124, 125, 136

Butler, J. M. (2005) *Forensic DNA Typing: Biology, Technology, and Genetics of STR Markers*. Academic Press, 2nd edn. 125, 129

Defays, D. (1977) An efficient algorithm for a complete link method. *The Computer Journal*, **4**, 364–366. 129

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, **39**, 1–38. 125

Excoffier, L. and Lischer, H. L. (2010) Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources*, **10**, 564–567. 132

Excoffier, L., Smouse, P. E. and Quattro, J. M. (1992) Analysis of molecular variance inferred from metric distances among dna haplotypes: Application to human mitochondrial dna restriction data. *Genetics*, **131**, 479–491. 124, 125, 132, 136

Hahsler, M., Buchta, C. and Hornik, K. (2012) *Infrastructure for seriation*. URL `http://CRAN.R-project.org/`. R package version 1.0-7. 129

Hahsler, M., Hornik, K. and Buchta, C. (2008) Getting things in order: An introduction to the R package seriation. *Journal of Statistical Software*, **25**, 1–34. 129

Lerski, G. J. (1996) *Historical dictionary of Poland, 966-1945*. Westport, Connecticut, USA: Greenwood. 136

R Development Core Team (2013) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org`. ISBN 3-900051-07-0. 124, 125, 129

Roewer, L., Croucher, P. J. P., Willuweit, S., Lu, T. T., Kayser, M., Lessig, R., de Knijff, P., Jobling, M. A., Tyler-Smith, C. and Krawczak, M. (2005) Signature of recent historical events in the European Y-chromosomal STR haplotype distribution. *Human Genetics*, **116**, 279–291. 124, 125, 128, 136

Roewer, L., Krawczak, M., Willuweit, S. *et al.* (2001) Online reference database of European Y-chromosomal short tandem repeat (STR) haplotypes. *Forensic Science International*, **2-3**, 106–113. 129

Schwarz, G. (1978) Estimating the Dimension of a Model. *Annals of Statistics*, **6**, 461–464. 125

Sørensen, T. (1948) A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. *Biologiske Skrifter*, **5**, 1–34. 129

Ward, Jr., J. H. (1963) Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, **48**, 236–244. 133

Willuweit, S. and Roewer, L. (2009) Y chromosome haplotype reference database (YHRD): Update. *Forensic Science International: Genetics*, **1**, 83–87. 129

Appendix A. Kullback-Leibler distance measure

Let $f(d;p)$ be the probability mass function of the discrete Laplace distribution. For the k'th locus together with subpopulation g and h, let

$$z_1 = \hat{y}_{gk}, \quad z_2 = \hat{y}_{hk}, \quad p_1 = \hat{p}_{gk} \quad \text{and} \quad p_2 = \hat{p}_{hk},$$

such that the distance from subpopulation g to h can be defined as

$$\begin{aligned}
\mathrm{KL}'_k(g,h) &= \sum_{d\in\mathbb{Z}} f\left(|d-z_1|;p_1\right)\log\left(\frac{f\left(|d-z_1|;p_1\right)}{f\left(|d-z_2|;p_2\right)}\right) \\
&= \sum_{d\in\mathbb{Z}} \left(\frac{1-p_1}{1+p_1}\right) p_1^{|d-z_1|} \log\left(\frac{\left(\frac{1-p_1}{1+p_1}\right)p_1^{|d-z_1|}}{\left(\frac{1-p_2}{1+p_2}\right)p_2^{|d-z_2|}}\right) \\
&= \left(\frac{1-p_1}{1+p_1}\right) \\
&\quad \times \sum_{d\in\mathbb{Z}} p_1^{|d-z_1|}\left\{|d-z_1|\log p_1 + \log\left(\frac{1-p_1}{1+p_1}\right) - |d-z_2|\log p_2 - \log\left(\frac{1-p_2}{1+p_2}\right)\right\} \\
&= \left(\frac{1-p_1}{1+p_1}\right)\left(\mathrm{KL}_k^{(1)}(g,h) + \mathrm{KL}_k^{(2)}(g,h) + \mathrm{KL}_k^{(3)}(g,h) + \mathrm{KL}_k^{(4)}(g,h)\right),
\end{aligned}$$

where

$$\mathrm{KL}_k^{(1)}(g,h) = \sum_{d\in\mathbb{Z}} p_1^{|d-z_1|}|d-z_1|\log p_1 = \log p_1 \sum_{d\in\mathbb{Z}} |d| p_1^{|d|} = 2\log p_1 \sum_{d=1}^{\infty} d\, p_1^d = \frac{2p_1\log p_1}{(p_1-1)^2}$$

$$\begin{aligned}
\mathrm{KL}_k^{(2)}(g,h) &= \sum_{d\in\mathbb{Z}} p_1^{|d-z_1|}\log\left(\frac{1-p_1}{1+p_1}\right) = \log\left(\frac{1-p_1}{1+p_1}\right)\sum_{d\in\mathbb{Z}} p_1^{|d|} = \log\left(\frac{1-p_1}{1+p_1}\right)\left(1+2\sum_{d=1}^{\infty} p_1^d\right) \\
&= \log\left(\frac{1-p_1}{1+p_1}\right)\left(1+\frac{2p_1}{1-p_1}\right) = \left(\frac{1+p_1}{1-p_1}\right)\log\left(\frac{1-p_1}{1+p_1}\right)
\end{aligned}$$

$$\begin{aligned}
\mathrm{KL}_k^{(3)}(g,h) &= -\sum_{d\in\mathbb{Z}} p_1^{|d-z_1|}|d-z_2|\log p_2 = -\log p_2 \sum_{d\in\mathbb{Z}} |d-z_2| p_1^{|d-z_1|} \\
&= -\log p_2 \sum_{d\in\mathbb{Z}} |d-z_2+z_1| p_1^{|d|} = -\log p_2 \sum_{d\in\mathbb{Z}} |z_2-z_1-d| p_1^{|d|}
\end{aligned}$$

$$\mathrm{KL}_k^{(4)}(g,h) = -\sum_{d\in\mathbb{Z}} p_1^{|d-z_1|}\log\left(\frac{1-p_2}{1+p_2}\right) = -\log\left(\frac{1-p_2}{1+p_2}\right)\sum_{d\in\mathbb{Z}} p_1^{|d|} = -\left(\frac{1+p_1}{1-p_1}\right)\log\left(\frac{1-p_2}{1+p_2}\right).$$

To evaluate $\mathrm{KL}_k^{(3)}(g,h)$, note that

$$\begin{aligned}
\sum_{d\in\mathbb{Z}} |z_2-z_1-d| p_1^{|d|} &= \sum_{d\in\mathbb{Z}} |-(z_2-z_1)+d| p_1^{|d|} \\
&= \sum_{-d\in\mathbb{Z}} |-(z_2-z_1)-d| p_1^{|d|}
\end{aligned}$$

$$= \sum_{-d\in\mathbb{Z}} |z_2 - z_1 + d| p_1^{|d|},$$

such that for $m = |z_2 - z_1| \geqslant 0$,

$$\begin{aligned}
\sum_{d\in\mathbb{Z}} |z_2 - z_1 - d| p_1^{|d|} &= \sum_{d\in\mathbb{Z}} |m - d| p_1^{|d|} \\
&= \sum_{d=-\infty}^{0} |m - d| p_1^{|d|} + \sum_{d=1}^{m} |m - d| p_1^{|d|} + \sum_{d=m+1}^{\infty} |m - d| p_1^{|d|} \\
&= \sum_{d=0}^{\infty} (m + d) p_1^d + \sum_{d=1}^{m} (m - d) p_1^d + \sum_{d=m+1}^{\infty} (d - m) p_1^d \\
&= \frac{(1 - p_1)m + p_1}{(p_1 - 1)^2} + \frac{p_1(p_1^m - m(p_1 - 1) - 1)}{(p_1 - 1)^2} + \frac{p^{m+1}}{(p_1 - 1)^2} \\
&= \frac{2p_1^{m+1} - m(p_1^2 - 1)}{(p_1 - 1)^2},
\end{aligned}$$

resulting in

$$\mathrm{KL}_k^{(3)}(g, h) = -\left(\frac{2p_1^{m+1} - m(p_1^2 - 1)}{(p_1 - 1)^2}\right) \log p_2.$$

Thus,

$$\begin{aligned}
\mathrm{KL}_k'(g, h) &= \left(\frac{1 - p_1}{1 + p_1}\right) \frac{2p_1 \log p_1}{(1 - p_1)^2} + \log\left(\frac{1 - p_1}{1 + p_1}\right) \\
&\quad - \left(\frac{1 - p_1}{1 + p_1}\right) \left(\frac{2p_1^{m+1} - m(p_1^2 - 1)}{(p_1 - 1)^2}\right) \log p_2 - \log\left(\frac{1 - p_2}{1 + p_2}\right) \\
&= \frac{2p_1 \log p_1}{1 - p_1^2} + \log\left(\frac{(1 - p_1)(1 + p_2)}{(1 + p_1)(1 - p_2)}\right) - \left(\frac{2p_1^{m+1} - m(p_1^2 - 1)}{1 - p_1^2}\right) \log p_2.
\end{aligned}$$

To make the distance symmetric, let

$$\mathrm{KL}_k(g, h) = \mathrm{KL}_k'(g, h) + \mathrm{KL}_k'(h, g).$$

Because mutations are assumed to happen independently across loci, we can sum the distances at each locus such that

$$\mathrm{KL}(g, h) = \sum_{k=1}^{r} \mathrm{KL}_k(g, h)$$

is the distance between subpopulation g and h.

Paper IX

Efficient iteratively reweighted least squares for weighted two-way analysis of variance

Author list Mikkel Meyer Andersen, *Aalborg University, Denmark*
Poul Svante Eriksen, *Aalborg University, Denmark*

Summary Weighted two-way analysis of variance with only main effects is a special case of a generalized linear model that can be heavily optimised by exploiting structure in the design matrix. The optimisation is obtained by calculating sufficient statistics directly without constructing a potentially huge design matrix.

Publication info This paper is in preparation.

1. Introduction

Andersen *et al.* (2013) describe how to make inference using the expectation-maximization (EM) algorithm by Dempster *et al.* (1977) in a mixture model with a multivariate (marginally independent) exponential family as components for a specific application (modelling a particular type of DNA profiles).

The standard way to make the maximization step – which is also what Andersen *et al.* (2013) suggested – is to make inference in a generalized linear model. Then, the number of rows in the design matrix will be $n \times c \times r$, where n is the number of individuals, c the number of components in the mixture and r the number of dimensions of the multivariate distribution, such that the number of rows is huge, which will lead to slow inference. However, the design matrix is very structured (the same block of $c \times r$ rows is repeated n times) and this can be exploited in making the maximization step much more both memory and CPU efficient as will be described in this paper. Here, the setup will be slightly more general than the one described by Andersen *et al.* (2013) as the design does not need to be balanced and the exponential family does not need to be on canonical form.

2. Model

Assume that we have repeated observations under two independent conditions $j \in \{1,2,\dots,c\}$ and $k \in \{1,2,\dots,r\}$ with a certain weight. The number of observations for conditions combination (j,k) is n_{jk}. Hence, the observations are d_{ijk} with weights w_{ijk} for $i \in \{1,2,\dots,n_{jk}\}$, $j \in \{1,2,\dots,c\}$ and $k \in \{1,2,\dots,r\}$. Assume that the observations are distributed according to an exponential family with link function g.

Below, it is assumed that $r \leqslant c$ in order to invert the smallest possible matrix (explained below). If $r > c$, the conditions should be interchanged as this leads to more optimal computations.

The Kronecker delta is defined by

$$\delta_{pq} = \begin{cases} 1 & \text{if } p = q \text{ and} \\ 0 & \text{otherwise.} \end{cases}$$

The inference of interest is to estimate the main effects in the two-way layout with no interaction. More specifically, let

$$\begin{aligned} \mu_{ijk} &= \mathbf{E}[d_{ijk}] \\ \tau_{ijk} &= \mathbf{Var}[d_{ijk}] = V(\mu_{ijk})/w_{ijk} \\ g(\mu_{ijk}) &= \sum_{p=1}^{c} \delta_{pj}\alpha_p + \sum_{q=1}^{r-1} \delta_{qk}\lambda_q \end{aligned}$$

for $j \in \{1,2,\dots,c\}$, $k \in \{1,2,\dots,r\}$ and $i \in \{1,2,\dots,n_{jk}\}$, where $V(\mu_{ijk})$ is the variance function of the exponential family. Note, that there is only $c + r - 1$ effects to ensure uniqueness of the effects.

Let $\vec{\beta} = (\alpha_1, \alpha_2, \ldots, \alpha_c, \lambda_1, \lambda_2, \ldots, \lambda_{r-1})^\top$ be the parameter vector. Now, let $\hat{\vec{\beta}}^{(0)}$ be the initial parameter vector. Then the iteratively reweighted least squares (IRLS) algorithm with design matrix X requires the following steps until convergence:

$$\begin{aligned}
\vec{z}^{(m+1)} &= X\hat{\vec{\beta}}^{(m)} \\
\mu_{ijk}^{(m+1)} &= g^{-1}\left(z_{ijk}^{(m+1)}\right) \\
\tau_{ijk}^{(m+1)} &= V\left(\mu_{ijk}^{(m+1)}\right) / w_{ijk} \\
W^{(m+1)} &= \operatorname{diag}\left(\left\{\left[\tau_{ijk}^{(m+1)}\left\{g'\left(\mu_{ijk}^{(m+1)}\right)\right\}^2\right]^{-1}\right\}_{ijk}\right) \\
y_{ijk}^{(m+1)} &= w_{ijk}\left(d_{ijk} - \mu_{ijk}^{(m+1)}\right) \\
\vec{y}^{\top} &= \left\{y_{ijk}^{(m+1)}\right\}_{ijk} \\
\hat{\beta}^{(m+1)} &= \hat{\beta}^{(m)} + \left(X^\top W^{(m+1)} X\right)^{-1} X^\top W^{(m+1)} \vec{y}^{(m+1)}.
\end{aligned}$$

2.1. *Optimising iterations*

Note, that to estimate $\hat{\beta}^{(m+1)}$, the quantity

$$\left(X^\top W^{(m+1)} X\right)^{-1} X^\top W^{(m+1)} \vec{y}^{(m+1)}$$

must be calculated. This can be done without explicitly constructing X as will now be described.

Let $\left(\vec{e}_p\right)_{ijk}$ denote element ijk of $\vec{e}_p$, which e.g. can be represented as a three dimensional array with dimensions $\left(\{n_{jk}\}_{jk}, c, r\right)$. The element ijk is given by

$$\left(\vec{e}_p\right)_{ijk} = \delta_{jp}.$$

This would correspond to the p'th column of the design matrix. Similarly, let

$$\left(\vec{f}_q\right)_{ijk} = \delta_{kq}.$$

This means that

$$\begin{aligned}
X &= (\vec{e}_1, \vec{e}_2, \ldots, \vec{e}_c, \vec{f}_1, \vec{f}_2, \ldots, \vec{f}_{r-1}) \\
X\vec{\beta} &= \sum_{p=1}^{c} \vec{e}_p \alpha_p + \sum_{q=1}^{r-1} \vec{f}_q \lambda_q.
\end{aligned}$$

Let $W^{(m+1)} = \operatorname{diag}\left(\left\{\psi_{ijk}^{(m+1)}\right\}_{ijk}\right)$ such that

$$\psi_{ijk}^{(m+1)} = \left[\tau_{ijk}^{(m+1)}\left\{g'\left(\mu_{ijk}^{(m+1)}\right)\right\}^2\right]^{-1}.$$

For ease of notation, drop the iteration number $m+1$ as all the following calculations are performed in the same iteration. Then, the elements of $X^\top WX$ are

$$\begin{aligned}
\left(\vec{e}_p\right)^\top W\vec{e}_j &= \delta_{pj}\sum_{ik}\psi_{ijk} = \delta_{pj}\psi_{+j+}\\
\left(\vec{f}_q\right)^\top W\vec{f}_k &= \delta_{qk}\sum_{ij}\psi_{ijk} = \delta_{qk}\psi_{++k}\\
\left(\vec{f}_q\right)^\top W\vec{e}_p &= \sum_{i}\psi_{ipq} = \psi_{+pq}.
\end{aligned}$$

Let

$$\begin{aligned}
D_c &= \operatorname{diag}\left(\{\psi_{+j+}\}_j\right)\\
D_{r-1} &= \operatorname{diag}\left(\{\psi_{++k}\}_k\right)\\
H &= \{\psi_{+jk}\}_{jk},
\end{aligned}$$

where H is a $c\times(r-1)$ matrix having $(\psi_{+1k},\psi_{+2k},\ldots,\psi_{+ck})^\top$ as the k'th column. Then,

$$X^\top W^{(m+1)}X = \begin{bmatrix} D_c & H\\ H^\top & D_{r-1}\end{bmatrix}.$$

According to Seber (1984, Appendix A3.1), the inverse of this is

$$\left(X^\top W^{(m+1)}X\right)^{-1} = \begin{bmatrix} D_c & H\\ H^\top & D_{r-1}\end{bmatrix}^{-1} = \begin{bmatrix} D_c^{-1}+FE^{-1}F^\top & -FE^{-1}\\ -E^{-1}F^\top & E^{-1}\end{bmatrix},$$

where

$$\begin{aligned}
E &= D_{r-1} - H^\top D_c^{-1}H\\
F &= D_c^{-1}H.
\end{aligned}$$

Here, the demanding operation is to find E^{-1} from the $(r-1)\times(r-1)$ matrix E. This is the reason that the conditions should be interchanged such that $r\leqslant c$, as mentioned previously.

2.2. *Optimised scheme*

Because of the very structured format of X, $\vec{z}^{(m+1)}$ will include repeated elements that do not contribute to the inference. The same is true for the values derived from $\vec{z}^{(m+1)}$. Now, redefine the quantities to not depend on i in order to obtain the optimised scheme with

$$\hat{\vec{\beta}}^{(0)} = \left(\hat{\alpha}_1^{(0)},\hat{\alpha}_2^{(0)},\ldots,\hat{\alpha}_c^{(0)},\hat{\lambda}_1^{(0)},\hat{\lambda}_2^{(0)},\ldots,\hat{\lambda}_{r-1}^{(0)}\right)^\top$$

as the initial parameter vector and iterate the following steps until convergence:

$$z_{jk}^{(m+1)} = \hat{\alpha}_j^{(m)} + \hat{\lambda}_k^{(m)}$$

$$
\begin{aligned}
\mu_{jk}^{(m+1)} &= g^{-1}\left(z_{jk}^{(m+1)}\right) \\
\tau_{ijk}^{(m+1)} &= V\left(\mu_{jk}^{(m+1)}\right) / w_{ijk} \\
\psi_{ijk}^{(m+1)} &= \left[\tau_{ijk}^{(m+1)}\left\{g'\left(\mu_{jk}^{(m+1)}\right)\right\}^2\right]^{-1} \\
D_c^{(m+1)} &= \operatorname{diag}\left(\left\{\psi_{+j+}^{(m+1)}\right\}_j\right) \\
D_{r-1}^{(m+1)} &= \operatorname{diag}\left(\left\{\psi_{++k}^{(m+1)}\right\}_k\right) \\
H^{(m+1)} &= \left\{\psi_{+jk}^{(m+1)}\right\}_{jk} \\
E^{(m+1)} &= D_{r-1}^{(m+1)} - \left(H^{(m+1)}\right)^\top \left(D_c^{(m+1)}\right)^{-1} H^{(m+1)} \\
F^{(m+1)} &= \left(D_c^{(m+1)}\right)^{-1} H^{(m+1)} \\
P^{(m+1)} &= \begin{bmatrix} D_c^{-1} + FE^{-1}F^\top & -FE^{-1} \\ -E^{-1}F^\top & E^{-1} \end{bmatrix} \\
a_j^{(m+1)} &= \sum_{ik} w_{ijk}\left(d_{ijk} - \mu_{jk}^{(m+1)}\right) \\
b_k^{(m+1)} &= \sum_{ij} w_{ijk}\left(d_{ijk} - \mu_{jk}^{(m+1)}\right) \\
\vec{\gamma}^{(m+1)} &= \left(a_1^{(m+1)}, a_2^{(m+1)}, \ldots, a_c^{(m+1)}, b_1^{(m+1)}, b_2^{(m+1)}, \ldots, b_{r-1}^{(m+1)}\right)^\top \\
\hat{\beta}^{(m+1)} &= \hat{\beta}^m + P^{(m+1)}\vec{\gamma}^{(m+1)}.
\end{aligned}
$$

An implementation of this scheme is provided in Appendix A.

3. Application in mixtures

As noted in the introduction, Andersen *et al.* (2013) describe a problem that can be solved more efficiently by using the described optimised IRLS. In this section, we describe the set-up. The problem is tackled by using a mixture of exponential families, where the component that each observation originates from is unknown. To deal with this, the EM algorithm by Dempster *et al.* (1977) is used to estimate the probability for originating from each component for each observation.

The observation, x, relative to a known location parameter, t, is assumed to be distributed according to a one-parameter exponential family on canonical form, such that

$$P(x;\theta) = a(\theta)h(x)\exp(\theta|x-t|),$$

where $a(\theta)$ is the normalisation factor.

Let $z_i = j$ denote that the i'th observation originates from mixture component j and let $v_{ij} = \mathbf{1}_{\{z_i=j\}}$ be the indicator function of this such that $v_{ij} = 1$ when $z_i = j$ and 0 otherwise. The full likelihood for n observations, c mixture components

and r dimensions is

$$L\left(\{\theta_{jk}\}_{jk}, \{v_{ij}\}_{jk}; \{x_{ik}\}_{ik}\right) \propto \prod_{i=1}^{n} \prod_{j=1}^{c} \prod_{k=1}^{r} \left[a(\theta_{jk}) \exp(\theta_{jk} | x_{ik} - t_{jk}|)\right]^{v_{ij}},$$

where $\theta_{jk} = \alpha_j + \lambda_k$.

As z_i, and hence v_{ij}, is assumed unknown, the full likelihood can be maximised by using e.g. the EM algorithm by Dempster *et al.* (1977). The E step consists of estimating

$$\hat{v}_{ij} = \mathbf{E}[v_{ij} \mid x_i]$$

given current estimates of $\{\theta_{jk}\}_{jk}$. The M step consists of estimating $\{\theta_{jk}\}_{jk}$ given $\{\hat{v}_{ij}\}_{ij}$. Normally, the two-way layout in the M step (main effects of j and k) would be estimated using a generalized linear model (e.g. using the `glm.fit` function in `R`) with $\{\hat{v}_{ij}\}_{ij}$ as weights as described by Wedel and DeSarbo (1995). The design matrix would then have $n \times c \times r$ rows. Instead, the optimised IRLS as described above, can used.

In the `R` (R Development Core Team, 2013) library `disclapmix` Andersen and Eriksen (2013), the optimised IRLS method described above is implemented as an alternative to the traditional `glm.fit`. This makes it possible to make inference for a database with 20,000 DNA profiles (n) of 20 loci (r) assuming more than 150 mixture components (c) equalling larger datasets as that obtained from a yet unpublished collaborative YHRD study of 23 Y-STRs in various populations (personal communication with Lutz Roewer and Michael Nothnagel). Using traditional generalized linear model (GLM) inference (e.g. via `glm.fit`), the design matrix, X, has 6×10^7 rows.

The German population of the yet unpublished collaborative YHRD study of $r = 23$ Y-STRs consists of $n = 1{,}690$ DNA profiles. Comparing the traditional GLM inference with the more efficient method described above yield speed improvements of almost 20 times for the optimal model for $c = 20$ (measured by the Bayesian Information Criterium (BIC) by Schwarz (1978)) and more for higher dimensional models. Even more can be gained by using the maximum relative change in the coefficient vector as stopping criterium instead of the deviance changes. See more details in Table 1. As seen, the speed-up increases with the dimension of the model. In practise, one could use the maximum relative change in the coefficient vector as stopping criterium until convergence and afterwards continue until the deviance criterium is met. In this way, the best from both can be utilised.

		Method		
		Efficient IRLS (coef)	Efficient IRLS (dev)	`glm.fit` (dev)
$c = 1$	Time	0.03 sec	0.06 sec	0.63 sec
	Speed-up	19 x	11 x	1 x
	Total time	0.07 sec	0.12 sec	1.27 sec
$c = 5$	Time	0.07 sec	0.25 sec	3.03 sec
	Speed-up	43 x	12 x	1 x
	Total time	3.38 sec	11.78 sec	145.59 sec
$c = 10$	Time	0.14 sec	0.51 sec	6.59 sec
	Speed-up	49 x	13 x	1 x
	Total time	2.70 sec	10.12 sec	131.84 sec
$c = 20$	Time	0.26 sec	1.00 sec	17.40 sec
	Speed-up	66 x	17 x	1 x
	Total time	11.33 sec	43.13 sec	748.08 sec
$c = 30$	Time	0.38 sec	1.49 sec	31.90 sec
	Speed-up	84 x	21 x	1 x
	Total time	24.32 sec	95.62 sec	2,041.64 sec
$c = 40$	Time	0.51 sec	1.99 sec	51.16 sec
	Speed-up	101 x	26 x	1 x
	Total time	43.14 sec	169.09 sec	4,348.99 sec
$c = 50$	Time	0.65 sec	2.50 sec	76.56 sec
	Speed-up	118 x	31 x	1 x
	Total time	69.14 sec	267.99 sec	8,192.30 sec

Table 1. Comparison study using $n = 1{,}690$ DNA profiles (with $r = 23$ Y-STR loci) from the German population. The time is the median time for a converged IRLS fit. Speed-up is the time compared to that of `glm.fit`. Total time is the median time for the EM algorithm to converge as described by Andersen *et al.* (2013), i.e. a complete model fit. Two different convergence criteria have been used: (dev) means that the deviance has been used as convergence criterium and (coef) means that maximum relative change in the coefficient vector has been used. The comparison was made on a desktop computer with an Intel® Core™ i7 CPU model 2600 running at 3.40GHz and 12 GB RAM. Measured by the Bayesian Information Criterium (BIC) by Schwarz (1978), 20 mixture components are optimal.

4. Bibliography

Andersen, M. M. and Eriksen, P. S. (2013) *disclapmix: Discrete Laplace mixture inference using the EM algorithm*. URL `http://CRAN.R-project.org/package=disclapmix`. R package version 1.2. 146

Andersen, M. M., Eriksen, P. S. and Morling, N. (2013) The discrete Laplace exponential family and estimation of Y-STR haplotype frequencies. *Journal of Theoretical Biology*, **329**, 39–51. 142, 145, 147

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, **39**, 1–38. 142, 145, 146

R Development Core Team (2013) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org`. ISBN 3-900051-07-0. 146

Schwarz, G. (1978) Estimating the Dimension of a Model. *Annals of Statistics*, **6**, 461–464. 146, 147

Seber, G. A. F. (1984) *Multivariate Observations*. Wiley. 144

Wedel, M. and DeSarbo, W. S. (1995) A Mixture Likelihood Approach for Generalized Linear Models. *Journal of Classification*, **12**, 21–55. 146

Appendix A. Implementation in R

```
> # d and w are arrays with dimension (n, c, r), where n, c and r are
    defined in the main text
>
> IRLS <- function(d, w, family, beta_start = NULL, verbose = !FALSE,
>                  eps = 1e-6, maxit = 25L,
>                  return_linear_predictors = FALSE) {
>
>   individuals <- dim(d)[1L]
>   clusters <- dim(d)[2L]
>   loci <- dim(d)[3L]
>
>   converged <- FALSE
>   lin_pred <- NULL
>   dev <- 0
>   devold <- Inf
>
>   # Beta initialisation
>   beta <- c(rep(-1, clusters), rep(0.5, loci - 1L))
>
>   if (!is.null(beta_start)) {
>     beta <- beta_start
>   }
>
>   for (iter in 1L:maxit) {
>     # Deviance
>     beta_dev <- c(beta, 0)
>     lin_pred <- rep(beta_dev[1L:clusters], each = individuals)
>     lin_pred <- lin_pred + rep(beta_dev[(clusters+1L):(clusters+loci)],
>       each = clusters * individuals)
>
>     mu_m <- family$linkinv(lin_pred)
>     dev <- sum(family$dev.resids(d, mu_m, w))
>
>     if (verbose == TRUE) {
>       cat("  IWLS iteration ", iter, ",
>         deviance = ", dev, "\n", sep = "")
>     }
>
>     if (abs(dev - devold)/(0.1 + abs(dev)) < eps) {
>       converged <- TRUE
>       break
>     }
>
>     devold <- dev
>
>     # Calculating sufficient statistics
>     beta_generic <- c(beta, 0)
>     lin_pred_generic <- outer(beta_generic[1L:clusters],
>       beta_generic[(clusters+1L):(clusters+loci)], "+")
>     mu_generic <- t(apply(lin_pred_generic, 1L, family$linkinv))
>
>     psi <- array(0, c(individuals, clusters, loci))
>
>     for (j in 1L:clusters) {
>       for (k in 1L:loci) {
>         psi[, j, k] <- ( w[, j, k] *
```

```
>           family$mu.eta(lin_pred_generic[j, k])^2 ) /
>             family$variance(mu_generic[j, k])
>       }
>     }
>
>     H <- array(NA, c(clusters, loci))
>     for (j in 1L:clusters) {
>       for (k in 1L:loci) {
>         H[j, k] <- sum(psi[, j, k])
>       }
>     }
>
>     Dr <- diag(colSums(H)[1L:(loci - 1L)], nrow = loci - 1L,
>       ncol = loci - 1L)
>     Dcinv <- diag(1 / rowSums(H), nrow = clusters, ncol = clusters)
>     H <- H[, 1L:(loci - 1L)]
>
>     E <- Dr - t(H) %*% Dcinv %*% H
>     Einv <- solve(E)
>     F <- Dcinv %*% H
>
>     P <- matrix(0, nrow = clusters + loci - 1L,
>       ncol = clusters + loci - 1L)
>     P[1L:clusters, 1L:clusters] <- Dcinv + F %*% Einv %*% t(F)
>     P[1L:clusters, (clusters + 1L):(clusters + loci - 1L)] <- -F %*% Einv
>     P[(clusters + 1L):(clusters + loci - 1L),
>       1L:clusters] <- -Einv %*% t(F)
>     P[(clusters + 1L):(clusters + loci - 1L),
>       (clusters + 1L):(clusters + loci - 1L)] <- Einv
>
>     d_mu_res <- array(0, c(individuals, clusters, loci))
>     for (i in 1L:individuals) {
>       d_mu_res[i, , ] <- d[i, , ] - mu_generic
>     }
>
>     a <- unlist(lapply(1L:clusters,
>       function(j) sum(w[,j,] * d_mu_res[,j,])))
>     b <- unlist(lapply(1L:(loci - 1L),
>       function(k) sum(w[,,k] * d_mu_res[,,k])))
>
>     gamma <- c(a, b)
>
>     beta_correction <- P %*% gamma
>     beta <- beta + beta_correction
>   }
>
>   coefficients <- as.numeric(beta)
>
>   if (!return_linear_predictors) {
>     lin_pred <- NULL
>   }
>
>   ans <- list(
>     coefficients = coefficients,
>     converged = converged,
>     deviance = dev,
>     linear.predictors = lin_pred
>   )
>
```

```
>   return(ans)
> }
>
> # Here follows an example for illustration only
> # (the dataset is too small to see any effect)
>
> # Contrasts specified such that
> # the same effects are fitted for both methods
> lmfit <- lm(breaks ~ tension + wool - 1, warpbreaks,
>   contrasts = list(
>     wool = contr.treatment(n = levels(warpbreaks$wool),
>                            base = length(levels(warpbreaks$wool)))))
>
> # Construct d and w such that r <= c
> l <- lapply(split(warpbreaks, warpbreaks$wool),
>   function(df) do.call(cbind,
>     lapply(split(df, df$tension), function(df2) df2$breaks)))
> d <- array(unlist(l), c(nrow(l[[1L]]), ncol(l[[1L]]), length(l)))
> w <- array(rep(1, length(d)), dim(d))
> irlsfit <- IRLS(d = d, w = w, family = gaussian(),
>   return_linear_predictors = TRUE)
>
> # Check the result
> coef(lmfit)
> irlsfit$coefficients
> irlsfit$coefficients - coef(lmfit)
> deviance(lmfit) - irlsfit$deviance
> sum((irlsfit$linear.predictors - predict(lmfit))^2)
```

Part 3

Epilogue

Conclusion

In this thesis, several models for lineage DNA markers have been presented. The work range from modelling errors introduced by chemicals and apparatus to population genetic work on how to estimate haplotype frequencies.

Most work was put into the discrete Laplace method, which is also reflected in the papers included. The main results in this thesis indicate that modelling of Y chromosomal short tandem repeat (Y-STR) haplotypes is done well by a finite mixture of discrete Laplace distributions ('the discrete Laplace method'). Both inference of haplotype frequencies and cluster analysis using this method (which has been implemented in publicly available software) yield state of the art results.

Future research

In this chapter, I will briefly describe the topics that I consider as the major parts of my future research. It is divided into extensions to existing work and new areas.

Appendix A. Extensions to existing work

1.1. *Validation of the discrete Laplace method*

Andersen *et al.* (2013) validated the discrete Laplace method based on populations following the Fisher-Wright model of evolution by Fisher (1922, 1930, 1958); Wright (1931); Ewens (2004) with assumptions of primarily neutral, single-step mutations of STRs (Ohta and Kimura, 1973). The performance of the discrete Laplace method should also be investigated for other types of populations and mutation models such as the logistic mutation model by Jochens *et al.* (2011).

1.2. *Robustness of the discrete Laplace method*

First, assume that we have a database with n DNA profiles. Next, assume that a biological trace containing a DNA profile, T, has been found at a crime scene and that a suspect has DNA profile S. Before calculating the match probability of S, S would have to be included in the database. An interesting question is: How much extra information is gained by adding S to the database?

This is a difficult question to answer, but one way to approach it is by considering the discrete Laplace method. Here, the match probabilty can be calculated in two situations: One based on the database without S and one based on the database with S included.

This has been done for each observation in 21 loci Y-STR haplotype databases (PowerPlex Y23 Y-STR haplotypes excluding DYS385a/b) from five populations (Danish, German, Italian, Spanish, Swedish) obtained from a yet unpublished collaborative YHRD study of 23 Y-STRs in various populations containing more than 18,000 haplotypes (personal communication with Lutz Roewer and Michael Nothnagel). The German database e.g. consists of 1,690 haplotypes. A discrete Laplace model was fitted for the entire database, and then 1,690 discrete Laplace models were fitted for each of the databases of size 1,689 obtained by removing each observation in turn. Similar inferences were performed with 15, 10 and 7 Y-STR loci.

In Figure 1, the comparison for the 7 Y-STR loci German population ($n = 1,690$) is shown. As seen, almost every point is on the straight line. This means that the Pearson correlation is high.

In Figure 2, the Pearson correlation for each population and number of loci is shown. Note, that for some databases and populations, the inclusion of the haplotype yields very different results. This is e.g. the case for the Swedish 10

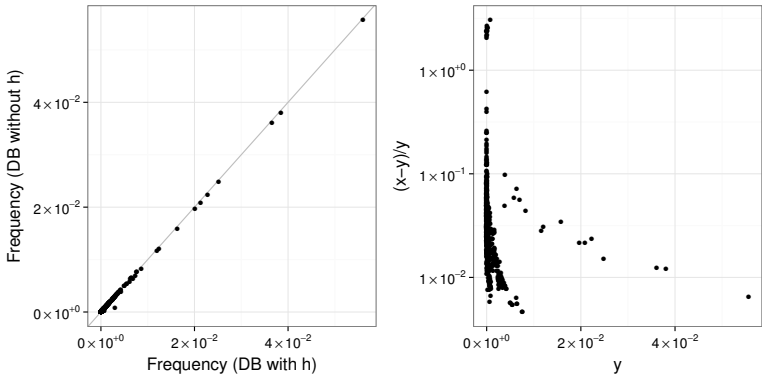


Figure 1. Comparison of estimated 7 loci Y-STR haplotype frequencies in the German population sample ($n = 1{,}690$) when including and not including a haplotype, h. Left: The straight line has an intercept of 0 and a slope of 1, such that it illustrates where the ordinate equals the abscissa. Right: x is Frequency (DB with h) and y is Frequency (DB without h). The straight line has an intercept of 0 and a slope of 0. As seen, $x - y > 0$, meaning that the frequency is higher when a profile is included in the database.

loci database ($n = 296$) shown in Figure 3. As seen, some haplotypes are severely underestimated when they are not included in the database.

Parts of this seem to be caused by the choice of the optimal number of clusters chosen by the discrete Laplace method, meaning that the number of clusters is not always robustly determined. If the same number of clusters is used for the fits without haplotype h as was chosen optimal for the entire database including h, the result shown in Figure 4 is obtained. Hence, not including a haplotype might cause a different number of optimal clusters.

Besides the optimal number of clusters chosen, the initial values of the central haplotypes of the clusters also play an important role, although the central haplotypes are allowed to changed during the inference as described by Andersen *et al.* (2013). This can be seen by choosing the initial central haplotypes in various ways and comparing the marginal BIC values of the resulting model fits. Such an analysis was done for the 10 loci Swedish database ($n = 296$). The method suggested by Andersen *et al.* (2013), based on experience, is partitioning around medoids (PAM) by Kaufman and Rousseeuw (1990). Besides this, three other methods were used. First, a principal component analysis (PCA) was made. From this, a k-means clustering was used to cluster the data (assigning each observation to a cluster). Then, the median in each dimension of the data points in each cluster was used as the initial central haplotypes. Second, a method by Kaufman and Rousseeuw (1990) similar to PAM called CLARA (that is based on simulation) was tried with 20 different random seeds. Third, observations were randomly chosen as central haplotypes. The results of this analysis are shown in Figure 5. As seen, some instances of CLARA gave better results than PAM, but this is not always the case, especially not for larger databases.

As demonstrated in this section, the likelihood function in the discrete Laplace method seems to have many local maxima, which sometimes can cause

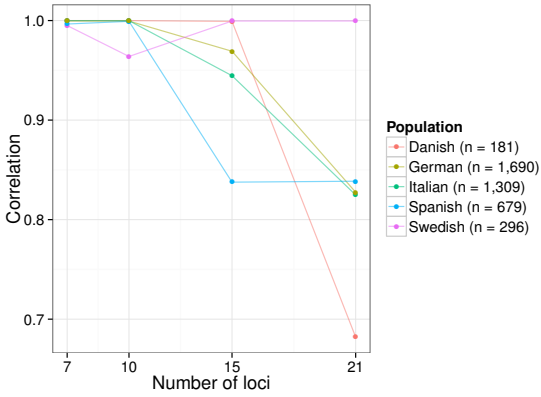


Figure 2. The Pearson correlation for each population and number of loci.

lack of robustness. This is an inherent problem due to the curse of dimensionality with such high dimensional data. However, this problem can most likely be solved for the discrete Laplace method. Hence, additional research in exploring the likelihood function is intended.

1.3. *Cluster analysis*

Cluster analysis using the discrete Laplace method of the data of a yet unpublished collaborative YHRD study of 23 Y-STRs in various populations including more than 18,000 haplotypes is work in progress (personal communication with Lutz Roewer and Michael Nothnagel). The results are to be compared with those obtained in paper VIII.

1.4. *Mixture analysis*

Because the discrete Laplace method can estimate frequencies of unobserved haplotypes, the method can be used for analysing Y-STR mixtures. One application is to make a deconvolution of a mixture to obtain the most probable individual profiles.

To assess how well such a deconvolution would work, a database of Y-STR profiles can be used. From this, two profiles, h_1 and h_2, can be drawn randomly and excluded from the database. Now, a discrete Laplace model can be estimated for the restricted database without h_1 and h_2. A mixture is then made from h_1 and h_2. Now, a deconvolution of this mixture can be made using the discrete Laplace model, e.g. by maximising the simultaneous probability of observing h_1 and h_2. This procedure must be repeated a certain number of times, e.g. 1,000 times. Similar analyses can be done for mixtures with three or more individuals.

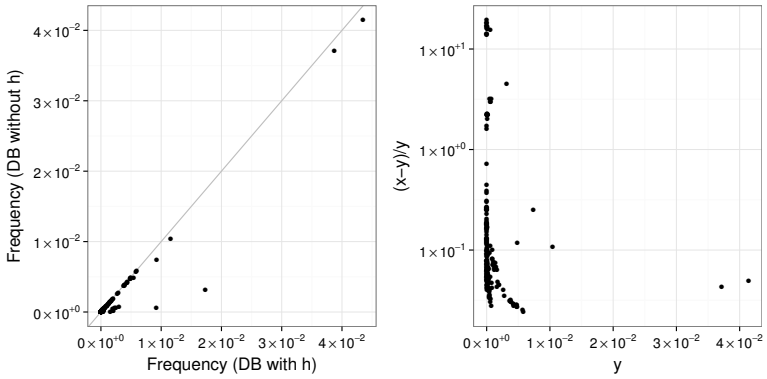


Figure 3. Comparison of estimated 10 loci Y-STR haplotype frequencies for the Swedish population sample ($n = 296$) when including and not including a haplotype, h. Left: The straight line has an intercept of 0 and a slope of 1, such that it illustrates where the ordinate equals the abscissa. Right: x is Frequency (DB with h) and y is Frequency (DB without h). The straight line has an intercept of 0 and a slope of 0. As seen, $x - y > 0$, meaning that the frequency is higher when a profile is included in the database.

Appendix B. New areas

2.1. *Population structure*

In this section, I briefly describe subjects related to population structure. I intend to do research in all these areas as they all have a huge impact on the use of lineage markers in forensic geneticists' everyday life of evidence interpretation.

Database collection

Evidential weight is calculated based on observed haplotypes and assuming certain population characteristics. Because the haplotypes of the entire population are not known, a sample from what is believed to be the population of interest is used. The way that a sample is assembled is essential for using it correctly. Normally, it is required that a sample consists of independent observations (a random sample) in order to analyse it using traditional statistical methods. This means that two closely related individuals both can be included in the sample simply by coincidence. And that is the way it should be.

If a random sample is taken and certain observations are excluded afterwards, for example due to assumed relationship (e.g. determined with autosomal STR analysis), then the sample is no longer a random sample and it is difficult or even impossible to use it for sound statistical analyses. Bodner *et al.* (2011) describe how to obtain 'better mtDNA population samples in forensic databases' by sampling unrelated individuals (sometimes this is refered to as 'sampling lineages') and their conclusion is:

> *The presence of maternally related donors in a "random" population sample has so far not been as thoroughly addressed in quality control as other aspects of mtDNA analysis and databasing. The simple*

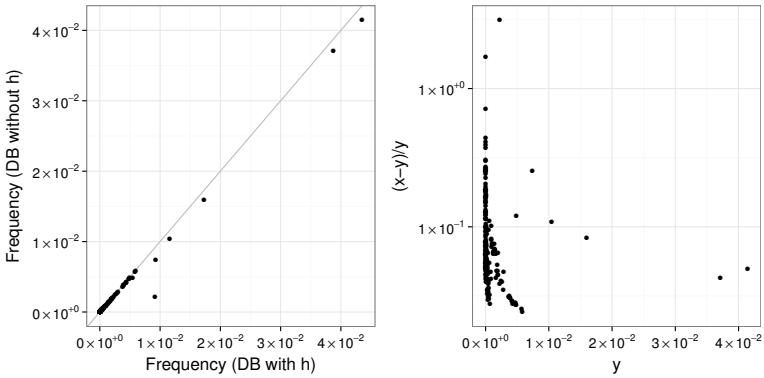


Figure 4. Comparison of estimated 10 loci Y-STR haplotype frequencies for the Swedish population sample ($n = 296$) when including and not including a haplotype, h. The number of clusters used was the same as that with the complete database. Left: The straight line has an intercept of 0 and a slope of 1, such that it illustrates where the ordinate equals the abscissa. Right: x is Frequency (DB with h) and y is Frequency (DB without h). The straight line has an intercept of 0 and a slope of 0. As seen, $x - y > 0$, meaning that the frequency is higher when a profile is included in the database.

> *practical approach presented here helps to detect the "clear and easy" cases of close maternal kinship between donors in a sample set: following the procedure described, these samples can be identified and subsequently excluded. If appreciated, this additional tool will contribute towards better random mtDNA population samples representative for their population, for the benefit of all research applying mtDNA as a genetic marker.*

Such a filtered sample with certain observations excluded may be usable for other analyses, but for evidential weight calculations, the haplotypes (and the number of times that they have been observed) is important information.

When a sample is filtered by excluding haplotypes from related individuals, the sample will not correctly reflect the frequency of the haplotypes as the haplotypes filtered will be underrepresented in the sample. This contradicts traditional statistical inference, where a sample must consist of independent observations.

To further emphasise the importance of having truly random samples: Almost all statistical methods assume random samples. If the sample is not random, the statistical results are not reliable.

Subpopulation correction

There has been some debate lately about incorporating knowledge about possible population structure when calculating the evidential weight of lineage markers, e.g. Buckleton *et al.* (2011).

It seems like there is still not consensus about how to calculate the evidential weight of lineage markers. It may have something to do with the interpretation of the standard defender's hypothesis stating:

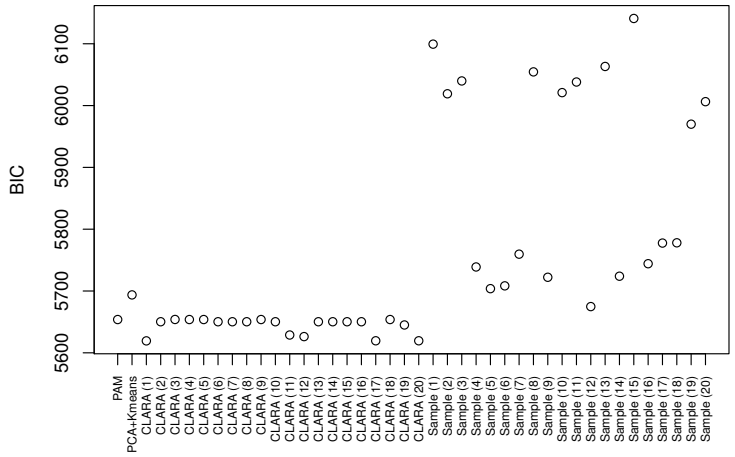


Figure 5. Comparison of the marginal BIC values of the resulting model fits when chosing the initial central haplotypes in various ways.

> *The probability that the suspect matches the haplotype found at the crime scene given that the suspect is unconnected to the crime.*

This is often translated to:

> *The probability that a random man's haplotype matches the haplotype found at the crime scene.*

Then one essential bit is specifying what is meant by a random man (e.g. what population, relationship to the offender, etc.). If it e.g. is believed that the offender originates from a certain part of a country, but only a country-wide sample is available, can the evidential weight be calculated using such a sample? And similarly for worldwide samples versus country specific samples.

Another consideration is if population structure corrections also should be made to results from methods that already model some population structure like the discrete Laplace method by Andersen *et al.* (2013).

Combining lineage markers and autosomal markers

Sometimes, both lineage markers and autosomal markers (or both Y chromosomal and mitochondrial DNA markers) are available. It is not obvious how an evidential weight from each of these should be combined to one evidential weight. Hence, methods for calculating the evidential weight using a combination of the different marker types have to be developed.

2.2. *Binary lineage DNA marker haplotypes*

The discrete Laplace method by Andersen *et al.* (2013) is a mixture model for Y-STR haplotypes. Each component consists of independent discrete Laplace distributions that model the STR alleles.

In principle, it should be possible to make a model for binary lineage DNA marker haplotypes such as mitochondrial DNA (mtDNA) and Y chromosomal single nucleotide polymorphism (Y-SNP) haplotypes by using a mixture of independent Bernoulli distributions. It requires a reference haplotype. The binary marker can then be whether the individual's marker is similar to that of the reference or not.

In NCBI dbSNP build 137 by Sherry *et al.* (2001), 1,246 single nucleotide substitutions were observed in the mtDNA. Of these, only 56 had substitutions of more than one nucleotide besides the one defined by the rCRS. Using these numbers, more than 90 % of the mtDNA substitutions can be thought of as a binary marker although it cannot be excluded that polymorphisms with three or four variants may be found. It would be interesting to use databases of mtDNA variations such as EMPOP (`http://www.empop.org/`) by Parson and Dür (2007) for estimating the fraction of mtDNA variations that can be assumed to be binary markers.

If more than one substitution is observed for a marker, the Bernoulli model requires one of at least two different approaches: (1) Disregard the marker or (2) group variations on a marker different from the reference in a single group. Both approaches mean that information is ignored or reduced. Hence, more advanced models must be considered, e.g. allowing for more than one variation at a position.

2.3. *Combining STR and SNP information*

The discrete Laplace method by Andersen *et al.* (2013) is a mixture model and so is the above mentioned Bernoulli mixture model for binary lineage DNA marker haplotypes. In principle, it should be possible to make a model for lineage haplotypes consisting of both STR markers and SNP markers. For example by combining the mixture models. It would be interesting to investigate this further. It might also help making the discrete Laplace model more robust in terms of clusters if including SNPs.

2.4. *Models for DNA sequences*

Another very interesting trend in genetics, including forensic genetics, is the use of second generating sequencing (SGS) also sometimes ambiguously referred to as next generation sequencing (NGS). SGS is a massively parallel sequencing technique that produces millions of reads (DNA fragments of up to 500 nucleotides). These reads can be mapped to a reference genome (or a part of it) such that consensus sequences of the sample sequenced can be constructed. A consensus sequence is typically made of up to hundreds of overlapping reads

per nucleotide. When sequencing haploid genomes, the aim is to obtain one consensus sequence and two when sequencing diploid genomes.

A consensus sequence may contain variations compared to the reference genome. These can either be actual variations or caused by errors in the sequencing process. If e.g. only one read out of hundred reads contains the variation, it is probably an error from the sequencing process. If all reads contain the variation, it is probably because the individual actually varies from the reference.

Variations that are interesting for forensic genetics include the traditional STR systems, where both length variations and complex repeats are relevant. The task of confirming a variation satisfactorily for forensic purposes is still not solved.

In this context, there are two major paths for analysing sequence data: 1) Determine when a variation is true and then use the variation as if it was confirmed. 2) Use all the reads for stating evidence about a possible variation.

Analogous to detecting STRs using electrophoresis, then 1) would be similar to using only the occurrence of alleles and neglect the peak heights and 2) would correspond to using all the peak heights, including stutters, noise etc. (sometimes referred to as continuous models).

Thus, second generating sequencing gives rise to a whole new era of data. Research in statistical methods for extracting high confidence information from SGS is planned.

Appendix C. Bibliography

Andersen, M. M., Eriksen, P. S. and Morling, N. (2013) The discrete Laplace exponential family and estimation of Y-STR haplotype frequencies. *Journal of Theoretical Biology*, **329**, 39–51. 157, 158, 162, 163

Bodner, M., Irwin, J., Coble, M. and Parson, W. (2011) Inspecting close maternal relatedness: towards better mtDNA population samples in forensic databases. *Forensic Science International: Genetics*, **5**, 138–141. 160

Buckleton, J., Krawczak, M. and Weir, B. (2011) The interpretation of lineage markers in forensic DNA testing. *Forensic Science International: Genetics*, **5**, 78–83. 161

Ewens, W. J. (2004) *Mathematical Population Genetics*. Springer-Verlag. 157

Fisher, R. A. (1922) On the Dominance Ratio. *Proc. Roy. Soc. Edin.*, **42**, 321–341. 157

Fisher, R. A. (1930) *The Genetical Theory of Natural Selection*. Oxford: Clarendon Press. 157

Fisher, R. A. (1958) *The Genetical Theory of Natural Selection*. New York: Dover, 2nd revised edn. 157

Jochens, A., Caliebe, A., Rösler, U. and Krawczak, M. (2011) Empirical Evaluation

Reveals Best Fit of a Logistic Mutation Model for Human Y-chromosomal Microsatellites. *Genetics*. 157

Kaufman, L. and Rousseeuw, P. J. (1990) *Finding groups in data: An introduction to cluster analysis*. Wiley. 158

Ohta, T. and Kimura, M. (1973) A Model of Mutation Appropriate to Estimate the Number of Electrophoretically Detectable Alleles in a Finite Population. *Genet. Res.*, **22**, 201–204. 157

Parson, W. and Dür, A. (2007) EMPOP – A forensic mtDNA database. *Forensic Science International: Genetics*, **1**, 88 – 92. 163

Sherry, S., Ward, M., Kholodov, M., Baker, J., Phan, L., Smigielski, E. and Sirotkin, K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*, **29**, 308–311. 163

Wright, S. (1931) Evolution in Mendelian populations. *Genetics*, **16**, 97–159. 157

Bibliography

Agresti, A. (2002) *Categorical Data Analysis*. Wiley, 2. edn.

Andersen, M. M. (2010) *Y-STR: Haplotype Frequency Estimation and Evidence Calculation*. Master's thesis, Aalborg University, Denmark.

Andersen, M. M., Caliebe, A., Jochens, A., Willuweit, S. and Krawczak, M. (2013a) Estimating trace-suspect match probabilities for singleton Y-STR haplotypes using coalescent theory. *Forensic Science International: Genetics*, **7**, 264–271.

Andersen, M. M. and Eriksen, P. S. (2012a) Efficient Forward Simulation of Fisher-Wright Populations with Stochastic Population Size and Neutral Single Step Mutations in Haplotypes. *Preprint, arXiv:1210.1773*.

Andersen, M. M. and Eriksen, P. S. (2012b) *fwsim: Fisher-Wright Population Simulation*. URL `http://CRAN.R-project.org/package=fwsim`. R package version 0.2-5.

Andersen, M. M. and Eriksen, P. S. (2013a) *disclap: Discrete Laplace Family*. URL `http://CRAN.R-project.org/package=disclap`. R package version 1.4.

Andersen, M. M. and Eriksen, P. S. (2013b) *disclapmix: Discrete Laplace mixture inference using the EM algorithm*. URL `http://CRAN.R-project.org/package=disclapmix`. R package version 1.2.

Andersen, M. M., Eriksen, P. S. and Morling, N. (2013b) A gentle introduction to the discrete Laplace method for estimating Y-STR haplotype frequencies. *Preprint, arXiv:1304.2129*.

Andersen, M. M., Eriksen, P. S. and Morling, N. (2013c) The discrete Laplace exponential family and estimation of Y-STR haplotype frequencies. *Journal of Theoretical Biology*, **329**, 39–51.

Andersen, M. M., Mogensen, H. S., Eriksen, P. S., Olofsson, J. K., Asplund, M. and Morling, N. (2013d) Estimating Y-STR allelic drop-out rates and adjusting for interlocus balances. *Forensic Science International: Genetics*, **7**, 327–336.

Andersen, M. M., Olofsson, J. K., Mogensen, H. S., Eriksen, P. S. and Morling, N. (2011) Estimating stutter rates for Y-STR alleles. *Forensic Science International: Genetics Supplement Series*, **3**, e192–e193.

Andersen, M. M. and Wilson, I. J. (2013) *rforensicbatwing: BATWING for calculating forensic trace-suspect match probabilities*. R package version 1.1.

Azzalini, A. (1996) *Statistical Inference – Based on the Likelihood*. Chapman & Hall.

Balding, D. J. and Nichols, R. A. (1994) DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and

single bands. *Forensic Sci. Int.*, **64**, 125–140.

Ballantyne, K. N. *et al.* (2010) Mutability of Y-Chromosomal Microsatellites: Rates, Characteristics, Molecular Bases, and Forensic Implications. *The American Journal of Human Genetics*, **87**, 341–353.

Bentley, J. L. (1975) Multidimensional Binary Search Trees Used for Associative Searching. *Communications of the ACM*, **18**, 509–517.

Bodner, M., Irwin, J., Coble, M. and Parson, W. (2011) Inspecting close maternal relatedness: towards better mtDNA population samples in forensic databases. *Forensic Science International: Genetics*, **5**, 138–141.

Breiman, L. (1996) Bagging predictors. *Machine Learning*, **24**, 123–140.

Brenner, C. H. (2010) Fundamental problem of forensic mathematics – The evidential value of a rare haplotype. *Forensic Science International: Genetics*, **4**, 281–291.

Brigham, E. O. (1988) *The fast Fourier transform and its applications*. Prentice Hall.

Brookes, C., Bright, J., Harbison, S. and Buckleton, J. (2012) Characterising stutter in forensic STR multiplexes. *Forensic Science International: Genetics*, **6**, 58–63.

Buckleton, J., Krawczak, M. and Weir, B. (2011) The interpretation of lineage markers in forensic DNA testing. *Forensic Science International: Genetics*, **5**, 78–83.

Budowle, B., Aranda, X. *et al.* (2008) Null allele sequence structure at the DYS448 locus and implications for profile interpretation. *International Journal of Legal Medicine*, **122**, 421–427.

Butler, J. M. (2001) *Forensic DNA Typing: Biology and Technology behind STR Markers*. Academic Press.

Butler, J. M. (2005) *Forensic DNA Typing: Biology, Technology, and Genetics of STR Markers*. Academic Press, 2nd edn.

Butler, J. M. (2006) Genetics and genomics of core short tandem repeat loci used in human identity testing. *Journal of Forensic Science*, **51**, 253–265.

Butler, J. M. (2010) *Fundamentals of Forensic DNA Typing*. Academic Press.

Butler, J. M. (2012) *Advanced Topics in Forensic DNA Typing: Methodology*. Academic Press.

Caliebe, A., Jochens, A., Krawczak, M. and Rösler, U. (2010) A Markov Chain Description of the Stepwise Mutation Model: Local and Global Behaviour of the Allele Process. *Journal of Theoretical Biology*, **266**, 336–342.

Cann, R., Stoneking, M. and Wilson, A. (1987) Mitochondrial DNA and human evolution. *Nature*, **325**, 31–36.

Cooley, J., Lewis, P. and Welch, P. (1969) The finite Fourier transform. *IEEE*

Trans. Audio Electroacoustics, **17**, 77–85.

Defays, D. (1977) An efficient algorithm for a complete link method. *The Computer Journal*, **4**, 364–366.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, **39**, 1–38.

Evett, I. W. and Weir, B. S. (1998) *Interpreting DNA Evidence*. Sinauer Associates.

Ewens, W. J. (2004) *Mathematical Population Genetics*. Springer-Verlag.

Excoffier, L. and Lischer, H. L. (2010) Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources*, **10**, 564–567.

Excoffier, L., Smouse, P. E. and Quattro, J. M. (1992) Analysis of molecular variance inferred from metric distances among dna haplotypes: Application to human mitochondrial dna restriction data. *Genetics*, **131**, 479–491.

Felsenstein, J. (2006) Accuracy of Coalescent Likelihood Estimates: Do We Need More Sites, More Sequences, or More Loci? *Mol. Biol. Evol.*, **23**, 691–700.

Fisher, R. A. (1922) On the Dominance Ratio. *Proc. Roy. Soc. Edin.*, **42**, 321–341.

Fisher, R. A. (1930) *The Genetical Theory of Natural Selection*. Oxford: Clarendon Press.

Fisher, R. A. (1958) *The Genetical Theory of Natural Selection*. New York: Dover, 2nd revised edn.

Gelman, A., Robers, G. O. and Gilks, W. R. (1996) Efficient Metropolis Jumping Rules. *Bayesian Statistics*, **5**, 599–607.

Gill, P., Brenner, C., Buckleton, J., Carracedo, A., Krawczak, M., Mayr, W., Morling, N., Prinz, M., Schneider, P. and Weir, B. (2006) DNA commission of the International Society of Forensic Genetics: Recommendations on the interpretation of mixtures. *Forensic Science International*, **160**, 90–101.

Gill, P., Brenner, C. *et al.* (2001) DNA Commission of the International Society of Forensic Genetics: Recommendations on forensic analysis using Y-chromosome STRs. *Forensic Science International*, **124**, 5–10.

Gill, P., Jeffreys, A. J. and Werrett, D. J. (1985) Forensic application of DNA fingerprints. *Nature*, **318**, 577–579.

Grün, B. and Leisch, F. (2008) FlexMix Version 2: Finite Mixtures with Concomitant Variables and Varying and Constant Parameters. *Journal of Statistical Software*, **28**.

Gusmao, L., Butler, J. *et al.* (2006) DNA Commission of the International Society of Forensic Genetics. DNA Commission of the International Society of Forensic Genetics (ISFG): an update of the recommendations on the use of Y-STRs in forensic analysis. *Forensic Science International*, **157**, 187–197.

Gusmao, L., Gonzalez-Neira, A., Alves, C., Lareu, M., Costa, S., Amorim, A. and Carracedo, A. (2002) Chimpanzee homologous of human Y specific STRs – A comparative study and a proposal for nomenclature. *Forensic Science International*, **126**, 129–136.

Hahsler, M., Buchta, C. and Hornik, K. (2012) *Infrastructure for seriation*. URL `http://CRAN.R-project.org/`. R package version 1.0-7.

Hahsler, M., Hornik, K. and Buchta, C. (2008) Getting things in order: An introduction to the R package seriation. *Journal of Statistical Software*, **25**, 1–34.

Hastings, W. K. (1970) Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika*, **57**, 97–109.

Hein, J., Schierup, M. H. and Wiuf, C. (2005) *Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory*. Oxford University Press.

Hosmer, D. W. and Lemeshow, S. (2000) *Applied Logistic Regression*. Wiley.

Hudson, R. R. (2001) Generating Samples Under a Wright–Fisher Neutral Model of Genetic Variation. *Bioinformatics*, **18**.

Inusah, S. and Kozubowski, T. J. (2006) A discrete analogue of the Laplace distribution. *Journal of Statistical Planning and Inference*, **136**, 1090–1102.

Jochens, A., Caliebe, A., Rösler, U. and Krawczak, M. (2011) Empirical Evaluation Reveals Best Fit of a Logistic Mutation Model for Human Y-chromosomal Microsatellites. *Genetics*.

Kaufman, L. and Rousseeuw, P. J. (1990) *Finding groups in data: An introduction to cluster analysis*. Wiley.

Kingman, J. F. C. (1982) The Coalescent. *Stochastic Processes and their Applications*, **13**, 235–248.

Krawczak, M. (2001) Forensic evaluation of Y-STR haplotype matches: a comment. *Forensic Science International*, **118**, 114–115.

Kullback, S. (1959) *Information theory and statistics*. John Wiley and Sons.

Kullback, S. and Leibler, R. A. (1951) On Information and Sufficiency. *Annals of Mathematical Statistics*, **22**, 79–86.

Leisch, F. (2004) FlexMix: A General Framework for Finite Mixture Models and Latent Class Regression in R. *Journal of Statistical Software*, **11**.

Lerski, G. J. (1996) *Historical dictionary of Poland, 966-1945*. Westport, Connecticut, USA: Greenwood.

Maechler, M., Rousseeuw, P., Struyf, A. and Hubert, M. (2005) Cluster analysis basics and extensions.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. and Teller, E. (1953) Equations of State Calculations by Fast Computing Machines. *Genetics*, **21**, 1087–1092.

Morling, N., Allen, R., Carracedo, A., Geada, H., Guidet, F., Hallenberg, C., Martin, W., Mayr, W., Olaisen, B., Pascali, V. and Schneider, P. (2002) Paternity Testing Commision of the International Society of Forensic Genetics: recommendations on genetic investigations in paternity cases. *Forensic Science International*, **129**, 148–157.

Ohta, T. and Kimura, M. (1973) A Model of Mutation Appropriate to Estimate the Number of Electrophoretically Detectable Alleles in a Finite Population. *Genet. Res.*, **22**, 201–204.

Olofsson, J. K., Andersen, M. M., Mogensen, H. S., Eriksen, P. S. and Morling, N. (2012) Sequence variants of allele 22 and 23 of DYS635 causing different stutter rates. *Forensic Science International: Genetics*, **6**, e161–e162. Letter to Editor.

Parson, W. and Dür, A. (2007) EMPOP – A forensic mtDNA database. *Forensic Science International: Genetics*, **1**, 88 – 92.

Persson, T. and Rootzen, H. (1977) Simple and Highly Efficient Estimators for a Type I Censored Normal Sample. *Biometrika*, **64**, 123–128.

Piazza, A., Mattiuz, P. and Ceppellini, R. (1969) [Combination of haplotypes of the HL-A system as a possible mechanism for gametic or zygotic selection]. *Haematologica*, **54**, 703–720. Article in Italian.

Prinz, M., Boll, K., Baum, H. and Shaler, B. (1997) Multiplexing of Y chromosome specific STRs and performance for mixed samples. *Forensic Science International*, **85**, 209–218.

R Development Core Team (2013) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

Robbins, H. E. (1968) Estimating the Total Probability of the Unobserved Outcomes of an Experiment. *The Annals of Mathematical Statistics*, **39**, 256–257.

Roewer, L. (2009) Y chromosome STR typing in crime casework. *Forensic Sci Med Pathol*, **5**, 77–84.

Roewer, L., Croucher, P. J. P., Willuweit, S., Lu, T. T., Kayser, M., Lessig, R., de Knijff, P., Jobling, M. A., Tyler-Smith, C. and Krawczak, M. (2005) Signature of recent historical events in the European Y-chromosomal STR haplotype distribution. *Human Genetics*, **116**, 279–291.

Roewer, L., Kayser, M., de Knijff, P. *et al.* (2000) A new method for the evaluation of matches in non-recombining genomes: application to Y-chromosomal short tandem repeat (STR) haplotypes in European males. *Forensic Science International*, **114**, 31–43.

Roewer, L., Krawczak, M., Willuweit, S. *et al.* (2001) Online reference database of European Y-chromosomal short tandem repeat (STR) haplotypes. *Forensic Science International*, **2-3**, 106–113.

Schwarz, G. (1978) Estimating the Dimension of a Model. *Annals of Statistics*, **6**, 461–464.

Seber, G. A. F. (1984) *Multivariate Observations*. Wiley.

Sherry, S., Ward, M., Kholodov, M., Baker, J., Phan, L., Smigielski, E. and Sirotkin, K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*, **29**, 308–311.

Sibille, I., Duverneuil, C. *et al.* (2002) Y-STR DNA amplification as biological evidence in sexually assaulted female victims with no cytological detection of spermatozoa. *Forensic Sci. Int.*, **125**, 212–216.

Skellam, J. G. (1946) The frequency distribution of the difference between two Poisson variates belonging to different populations. *Journal of Royal Statistical Society: Series A*, **109**, 296.

Sullivan, K., Hopgood, R., Lang, B. and Gill, P. (1991) Automated amplification and sequencing of human mitochondrial DNA. *Electrophoresis*, **12**, 17–21.

Sørensen, T. (1948) A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. *Biologiske Skrifter*, **5**, 1–34.

Titterington, D. M., Smith, A. F. M. and Makov, U. E. (1987) *Statistical Analysis of Finite Mixture Distributions*. Wiley.

Tvedebrink, T., Eriksen, P. S., Asplund, M., Mogensen, H. S. and Morling, N. (2011a) Allelic drop-out probabilities estimated by logistic regression - Further considerations and practical implementation. *Forensic Science International: Genetics*, **6**, 263–267.

Tvedebrink, T., Eriksen, P. S., Mogensen, H. S. and Morling, N. (2009) Estimating the probability of allelic drop-out of STR alleles in forensic genetics. *Forensic Science International: Genetics*, **3**, 222–226.

Tvedebrink, T., Eriksen, P. S., Mogensen, H. S. and Morling, N. (2011b) Statistical model for degraded DNA samples and adjusted probabilities for allelic drop-out. *Forensic Science International: Genetics*, **6**, 97–101.

Urquhart, A., Kimpton, C. P., Downes, T. J. and Gill, P. (1994) Variation in short tandem repeat sequences - a survey of twelve microsatellite loci for use as forensic identification markers. *Int. J. Med.*, **107**, 13–20.

Ward, Jr., J. H. (1963) Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, **48**, 236–244.

Wedel, M. and DeSarbo, W. S. (1995) A Mixture Likelihood Approach for Generalized Linear Models. *Journal of Classification*, **12**, 21–55.

Willuweit, S., Caliebe, A., Andersen, M. M. and Roewer, L. (2011) Y-STR Frequency Surveying Method: A critical reappraisal. *Forensic Science International: Genetics*, **5**, 84–90.

Willuweit, S. and Roewer, L. (2009) Y chromosome haplotype reference database

(YHRD): Update. *Forensic Science International: Genetics*, **1**, 83–87.

Wilson, I. J. and Balding, D. J. (1998) Genealogical Inference From Microsatellite Data. *Genetics*, **150**, 499–510.

Wilson, I. J., Weale, M. E. and Balding, D. J. (2003) Inferences from DNA Data: Population Histories, Evolutionary Processes and Forensic Match Probabilities. *Journal of Royal Statistical Society Series A*, **166**, 155–201.

Wright, S. (1931) Evolution in Mendelian populations. *Genetics*, **16**, 97–159.